Technical Standards for IEA Studies

Edited by

Michael O Martin

Keith Rust

Raymond J Adams

Contributors

Nancy Caldwell

Pierre Foy

Dirk Hastedt

Michael O Martin

Ina V S Mullis

International Association for the Evaluation of Educational Achievement

Technical Standards for IEA Studies

Edited by
Michael O Martin
Keith Rust
Raymond J Adams

Contributors
Nancy Caldwell
Pierre Foy
Michael O Martin
Ina V S Mullis
Heiko Sibberns

June 1999

International Association for the
Evaluation of Educational Achievement

ISBN 90 5166 7191

Copies of *Technical Standards for IEA Studies* can be obtained from:
IEA Secretariat
Herengracht 487
1017 BT Amsterdam
The Netherlands
Telephone +31 20 625 3625
Fax +31 20 420 7136
Email: Department@IEA.nl

Text edited by: Paula Wagemaker Editorial Services, Christchurch, New Zealand

Designed and desktop published by: Becky Bliss Design and Production, Wellington, New Zealand

Printed by: Eburon Publishers, Delft, Netherlands

Contents

Introduction

The International Association for the Evaluation of Educational Achievement (IEA) seeks to improve education through study of student achievement and the factors associated with it in educational systems around the world. In furtherance of this aim, IEA studies have made extensive, though not exclusive, use of sample survey methodology. While IEA studies have pioneered the application of sophisticated analytic techniques to the international comparative study of student achievement, and have always aspired to the highest methodological standards, there has not, until now, been any attempt to formulate these standards explicitly or to organise them in a systematic fashion.

This document was commissioned by the IEA Secretariat with the support of the United States National Center for Education Statistics (NCES) as part of an initiative to develop a set of technical standards that could be used as guides to good practice by future IEA studies. It has been prepared under the aegis of the IEA Technical Advisory Group. An early draft benefited from reviews and comment from the staff of the IEA Civics, Pre-Primary and SITES studies.

Work on the IEA Standards began with a review of several fairly recent and accessible standards documents. Among those consulted were:

- the *Standards for Education Data Collection and Reporting* (SEDCAR), produced for NCES by Westat (1991)
- the *NCES Statistical Standards* (1992)
- the *ETS Standards for Quality and Fairness* (Educational Testing Service, 1987)
- the *Standards for Educational and Psychological Testing* (American Psychological Association, 1985).

These documents were produced following long consultative processes involving many contributors and reviewers, and encapsulate much accumulated wisdom from decades of survey research and educational practice. As such, they contain much that usefully could be adapted to the IEA situation. IEA studies have their own unique demands in terms of the situation in which they operate and the knowledge and skills that they require, and a set of standards for such studies must take these situational factors as their point of departure.

On the basis of the review of existing material, we concluded that existing standards did not apply exactly to the unique situation in which IEA operates. We decided, therefore, to build a framework that incorporates the central characteristics of IEA studies, and to use this framework to develop a set of standards. While the framework would correspond to the structure of a typical IEA study, many of the tasks that are components of such a study are characteristic of investigations generally in social research; standards for the conduct of which have evolved over many years. Accordingly, we have adapted existing standards where they seemed to fit the purpose, as well as deriving new standards wherever the need arose.

Purpose and Scope of Technical Standards

The primary purpose behind delineating technical standards is to provide a ready yardstick against which IEA studies may be assessed.

The primary purpose behind delineating technical standards is to encapsulate the essentials of good practice in the field in which IEA conducts its studies, so as to provide a ready yardstick against which studies conducted under the auspices of IEA may be assessed. Although the standards are primarily criteria against which studies can be evaluated, they will also serve as guidelines for researchers involved in designing and implementing such studies. Studies will be judged in the light of these standards by the IEA General Assembly, perhaps on the recommendation of the IEA Technical Advisory Committee. Studies also will be judged against these standards by interested parties outside the IEA network who wish to make use of the results of the study or to use the products (tests, questionnaires, etc.) developed as part of a study.

There is some variation in how the published collections use the term 'standard'. In some, the term means a yardstick against which practices may be judged, without any indication of the degree of compliance required. For example, the SEDCAR standards use the term in this manner. The SEDCAR standards are presented as guidelines and examples of best practice, but do not specify limits as to what must be done, or supply criteria that must be met before studies are deemed acceptable. In contrast, the NCES Statistical Standards are more prescriptive, are very specific in terms of what is acceptable practice and go into considerable detail in some areas. The IEA standards in this present document follow the SEDCAR model.

Any set of standards for future studies must concentrate on assessing the technical and methodological excellence of those studies while not stifling innovation by circumscribing too closely what may be attempted.

IEA studies have in common a focus on cross-national comparisons of student achievement in school subjects, but beyond that individual studies have been allowed considerable latitude in choice of methodology, populations to be studied, data collection procedures, and analytic approaches. Ever since its inception in the late 1950s, the IEA has striven to use the latest and best methodology and technology in its projects. As methodology and technology have evolved, so have the conduct of IEA studies. Along with this evolution in methodology has come a greater sensitivity to the threats to the validity of studies caused by weaknesses in design and implementation, and a gradual demand for higher and more explicit standards. The search for ever higher and more explicit standards, while laudable, carries with it the danger of making unrealistic demands on projects. It is important that any set of standards for future studies concentrates on assuring the technical and

methodological excellence of those studies while not stifling innovation by circumscribing too closely what may be attempted.

We have chosen therefore to express the standards in a way that stresses the use of the most up to date methods, techniques and practices in each aspect of an IEA study, rather than take an approach that would have focussed attention on the prescription of specific tools and methods. Furthermore, we have not specified arbitrary target values for quantities such as response rates, test reliabilities and the like. Instead, we have stressed the value of setting goals for such values that will provide credibility for the project within the context of the specific study. We have also described practices that should lead to the maximisation of such values.

We have also chosen to focus on the most practical issues that must be addressed in conducting a study of high quality, and have emphasised planning, preparation, training and quality control at all stages of the enterprise. We did not attempt to provide a tutorial on the many advanced techniques in instrument development and psychometric scaling, sampling and data analysis, etc, since many fine texts already exist in these areas, and much valuable work has already been accomplished within the IEA community. The topics addressed in this volume are rarely discussed in a coherent fashion and are often taken for granted in standard texts.

Characteristics of IEA Studies

Any attempt to develop standards for IEA studies must take into consideration the co-operative nature of the organisation, and in particular its unique project structure and management style. *IEA studies operate on essentially two levels, national and international,* and any set of technical standards should accommodate this reality.

At the *international level*, studies are funded, planned and co-ordinated. An *international co-ordinating centre* (ICC)[1] is established to assume responsibility for all aspects of project design and implementation. Data collection instruments are developed, and data collection procedures, including quality assurance activities, are devised and documented. Data processing activities at the international level focus on checking the quality of the data and conducting analyses for international reports. The production of international reports and the dissemination of results are also the responsibility of the international project management.

At the *national level*, IEA *national centres* are responsible for funding and staffing IEA projects, and for appointing a *national research co-ordinator* and *national advisory committees* as appropriate. The national research co-ordinator is responsible for conducting the study within the participating system in accordance with the parameters of the study as agreed internationally. In particular, the national research co-ordinator is responsible for:

- developing survey population definitions and sampling plans that meet international specifications
- adapting and translating data collection instruments and procedural manuals
- implementing an acceptable data collection plan
- providing the resulting national data to the international project management in a standard international format
- conducting national analyses
- producing national reports as required by national authorities.

1 Both the terms 'international co-ordinating centre' and 'international study centre' have been used in IEA studies. This document uses the former for consistency with other published IEA information.

The overarching goal is to learn more about the factors that influence student attitudes and achievement.

IEA studies have traditionally focused on the output of educational systems, that is, the attitudes and educational achievements of students, and attempted to relate these outputs to the inputs that were antecedent to them. The overarching goal is to learn more about the factors that influence student attitudes and achievement, and which can be manipulated to bring about improvements in attitudes and achievement, or efficiencies in the educational enterprise.

The arena in which IEA has operated is the educational system, particularly the primary and secondary sectors. IEA has not paid attention to the post-secondary sector to date. The primary and secondary educational sectors are highly structured arenas, with students organised into classrooms and grouped according to grade levels within schools as an almost universal model. The form of the structure varies from country to country, and it is this variation across countries that IEA aims to study and relate to international variation in achievement. There is greater variation in instructional practices, educational facilities and student attitudes and achievement across a range of countries than within any one country, and so the study of such factors across countries can provide insights that would be unattainable from the study of a single country.

There is greater variation in instructional practices, educational facilities and student attitudes and achievement across a range of countries than within any one country. The study of such factors across countries can provide insights unattainable from the study of a single country.

IEA studies have a strong empirical basis, and have relied mainly on cross-sectional and longitudinal non-experimental designs, with data collection through sample survey methods. Studies make use of qualitative methods such as case studies and observational techniques from time to time as appropriate, but the main thrust of IEA studies has been to bring a strong quantitative orientation to the description and analysis of large-scale survey data.

The heart of most IEA studies to date has been a cross-national sample survey of student achievement in one or more school subjects. Mathematics, science and reading are the subjects that have received the greatest attention, although information technology and civic education have also been studied more than once. Student achievement has been measured by administering objective tests (usually tests of achievement that have been developed as part of the study) to samples of students that have been selected to be representative of national populations. Information about the students' backgrounds, attitudes and interests has been collected by means of self-report questionnaires administered to the students at the same time as the achievement tests. Information about the schools that the students attend has been collected by means of a school questionnaire addressed to the school principal. Information about classroom practices and teachers' experiences and attitudes has been collected usually by means of a questionnaire to the

classroom teacher. Information about the education system (e.g., the structure of the system, national and regional policies, locus of decision-making) has been amassed through questionnaires completed by national research co-ordinators.

IEA studies in the past have had certain characteristic features, and many of these features are likely to be characteristic of future studies as well. These features are summarised below.

- The measurement of student achievement in school subjects is a fundamental objective. Often the development of the instrument for measuring student achievement is a major component of the study. Collateral information usually is collected by means of questionnaires from the students, their teachers and the school principals at about the same time. The development of such questionnaires usually is also a major component of the study.
- Data are collected by means of sample surveys conducted in the school setting. Data-collection methods are designed to capitalise on school organisation. Sampling plans adopt multi-stage designs with schools and classrooms as stages in the sampling process, and school personnel are often called upon to act as data-providers and data-collection agents. Educational systems are highly structured settings, and IEA researchers have accumulated a high level of experience and expertise in working through such environments. The development of strategies for collecting internationally comparable data in school settings is an IEA specialty.

The development of strategies for collecting internationally comparable data in school settings is an IEA specialty.

- Data are collected simultaneously in a large number of countries by national teams using internationally agreed-upon instruments and following internationally agreed-upon procedures. The adaptation of the international instruments and procedures to national conditions is performed locally by the national team. National teams are free to add extra cognitive or questionnaire items as national options, provided these options do not compromise the international instruments. The national team is responsible for quality control of all national activities.
- The study is managed and co-ordinated by an international team, which develops instruments and procedures through a co-operative process. The international team is responsible for quality assurance aspects of the overall project, and for ensuring that survey instruments and survey procedures used in each country conform to the international standard. The international team conducts analyses and writes international reports.

Structure of the Standards

The future direction of IEA studies is not, of course, completely constrained by the experience of the past. However, the afore-mentioned features of past IEA studies are likely to retain central importance in any IEA studies in the immediate future. Accordingly, the IEA decided to begin the process of developing technical standards in the following four areas:

1. The design, management, operation and quality control of international studies.
2. The construction of instruments for measuring student achievement and questionnaires for collecting background information from students, teachers and schools.
3. Data collection by means of sample surveys in the school setting.
4. Data processing, analysis and reporting.

While the purpose of the standards is to help ensure that every facet of an IEA study is conducted in an optimal manner, there will undoubtedly be areas where national systems vary in their standards and norms of practice. For this reason, the IEA technical standards focus primarily on the design and management of studies at the international level, and on those aspects of the national implementation of studies that are necessary to ensure the timely collection of internationally comparable data of acceptable quality. Accordingly, many of the standards in this document are of primary relevance to staff of the international co-ordinating centre, whereas others apply to national staff. Each standard identifies the intended audience, as necessary.

IEA technical standards focus primarily on the design and management of studies at the international level, and on those aspects of the national implementation of studies that are necessary to ensure the timely collection of internationally comparable data of acceptable quality.

We have grouped the standards into the four areas described above, and within each area have specified six to nine standards. For each standard, we have provided a purpose, along with a statement that describes the standard. We have also included a set of guidelines to be followed so that the standard will be met, along with (for some standards) a checklist to help monitor the implementation of the standard.

Standards have been developed in the following areas:

Standards for designing, managing and implementing IEA studies

- Initial planning of an IEA study
- Choosing an international co-ordinating centre
- Formulating and refining study questions
- Designing an IEA study
- Developing a sampling plan
- Choosing data collection methods
- Developing a quality assurance programme
- Preparing an analysis plan.

Standards for developing data collection instruments

- Developing assessment frameworks and conceptual models
- Developing specifications for tests and questionnaires
- Test development
- Questionnaire development
- Translations and verifying translations
- Field testing data collection instruments and procedures.

Standards for data collection and processing

- Drawing a sample
- Planning for data collection
- Selecting and training data collection staff
- Minimising response burden and non-response
- Implementing data collection quality control procedures
- Documenting national data collection
- Planning for data preparation and processing
- Processing and checking data
- Documenting data-processing activities.

Standards for analysing data and reporting results

- Developing sampling weights
- Reporting sampling and non-sampling errors
- Validating constructs and scales for analysis
- Presenting study findings
- Reviewing the primary reports of study findings
- Releasing data
- Preparing technical reports and documentation.

The Standards

Designing, Managing and Implementing IEA Studies

Introduction

IEA studies are complex and costly endeavours. It therefore is important that great attention is paid at the outset to the specification of the aims and objectives of the study. Time and energy spent in the planning stages on establishing consensus as to study aims, objectives, methods and standards will be repaid many times over in terms of efficiency of operations during the lifetime of the study. The initial plan for the study should be a concise summary of the objectives, methods, schedule, costs and benefits of the projected study. It should contain sufficient information to enable intending funding agencies and participants to evaluate its scope and its projected costs and benefits.

The initial plan for the study should be a concise summary of the objectives, methods, schedule, costs and benefits of the projected study.

Given that international studies stand or fall on the quality of the international direction and co-ordination, an effective international co-ordinating centre is essential to the success of any IEA study. The centre must bring together individuals with the skills and experience necessary to master the technical and substantive aspects of the study and the management and communication abilities necessary to achieve and maintain international consensus while reaching project milestones in a timely manner.

Once the study gets under way, the elements of the initial plan must be expanded and developed as detailed guides for the implementation of the main tasks of the study. In particular, the study questions should be formulated and refined so that an efficient and economical design can be developed. Satisfactory implementation of the design will require the development of a sampling plan which specifies the sampling procedures that are appropriate for the study, and the standards for sampling precision that are required. Since the administration of tests and questionnaires is usually a central feature of IEA studies, an efficient scheme for collecting the data under standardised conditions is a central requirement. A comprehensive quality assurance program that monitors the data collection and records compliance with standard procedures is also a prerequisite for a modern IEA study.

A comprehensive quality assurance program that monitors the data collection and records compliance with standard procedures is a prerequisite for a modern IEA study.

An analysis plan that spells out how the data will be analysed so as to achieve the study aims is an essential component of the planning process. As well as providing assurance that the study aims can indeed be met with the data that are to be collected, a good analysis plan helps focus on the essentials of the study and reduces the temptation to engage in data-gathering that does not have a clear purpose.

Standard for Initial Planning of an IEA Study

Purpose:

To provide guidance in the initial design and planning of an IEA study.

Standard:

The initial plan for an IEA study should contain sufficient information to enable intending participants to evaluate its scope, projected costs and benefits. The plan should include the rationale for the study and a description of the goals and objectives, study methodology, target populations, analytic objectives and methods, implementation schedule, and preliminary cost and staffing estimates.

Guidelines:

To meet this standard, the plan should include the following:

a) *A justification of the study* in terms of the issues and questions to be addressed, stated in terms of the specific goals and objectives.

b) *A preliminary study design* that describes the target populations, sample design, instrument development, data collection methods, and the methodological issues to be resolved. The design should include requirements for acceptable population coverage and response rates.

c) *A clear distinction* between those components of the study that are compulsory for all participants and those components that may be excluded or modified at the discretion of a participant. Those areas where participants may exercise national options to collect information specific to their own systems should be identified.

d) *A preliminary analysis plan* that identifies analysis issues, objectives, major variables and proposed statistical techniques.

e) *A preliminary publication and dissemination plan* that identifies proposed major publications and their target audiences.

f) *A preliminary time schedule* for all the major project tasks over the complete study cycle, from planning to final publications. This schedule should indicate clearly which tasks are the responsibility of the international project management, which are the responsibility of the national research co-ordinators, and which are a joint responsibility.

g) *Preliminary international cost estimates*, broken down by major project task, for staff (in terms of person/months or weeks) and for all other costs.

Standard for Choosing an International Co-ordinating Centre

Purpose:

To ensure that the international co-ordinating centre is properly staffed, equipped and funded to carry out its function of co-ordinating all aspects of study design, development, implementation, analysis and reporting.

Standard:

As the international management and co-ordination centre for the study, the ICC should have access to staff with the technical and managerial competence and experience necessary to conduct the study successfully. In particular, staff will require the technical expertise to design and monitor the implementation of the study, as well as the managerial competence to plan and maintain realistic timelines and schedules, and to maintain close communication between the study partners, particularly national research co-ordinators, advisory committees and funding agencies.

The ICC should possess the infrastructure necessary to implement the study effectively, including accommodation, support staff, data processing equipment, and communications facilities (including electronic mail, fax and telephone). To maintain its operational readiness, the ICC should have access to a secure source of funds for the life of the study.

Guidelines:

The size and level of staffing of the international co-ordinating centre will depend on the scope and the level of funding for the study. For a modest study with limited funding, a small staff may be adequate, provided that the senior members possess the necessary technical and managerial competence to co-ordinate the study. More ambitious studies, with substantial funding and consequently higher profiles, require higher staffing levels, and often more specialised staff.

To meet the challenge of co-ordinating an IEA study, an ICC should have the following:

a) *Qualified staff*. At a minimum, the senior staff should together have the technical knowledge and expertise to understand all aspects of the design and operation of the study. They should also have the managerial competence and experience to set and maintain schedules, hire and motivate staff, develop and manage budgets, disseminate information about the study and maintain communications between study partners. A good knowledge of the subject matter under study (mathematics, science, language, etc.)

on the part of senior staff is also necessary, but such knowledge is no substitute for expertise and experience of survey research in a school setting.

The primary function of the ICC is to ensure that the study is implemented to the highest technical standards, and this requires senior staff skilled in the methods of survey research. IEA studies typically also require staff with a high level of competence in information technology, particularly desktop publishing and communications technology, and in electronic data processing.

Staff should also possess a high degree of fluency in the language of the study (to date, this has been English), so that all documents produced by the ICC meet the highest semantic and syntactic standards. It is desirable that senior staff be familiar with a second language, so that they will have an appreciation of the difficulties of language translation.

b) *Sound infrastructure*. The ICC requires suitable office accommodation, with adequate working and meeting space, and ready access to modern information technology, including computer workstations configured for word processing, spreadsheet work, database construction, data processing and analysis, and electronic mail. Other necessary office equipment includes photocopiers, fax machines, telephones and reliable postage or courier services.

c) *Secure funding*. If the ICC is to function effectively and attract and retain suitable staff, adequate funds must be secured at an early stage to support the operation of the ICC for the life of the study.

Standard for Formulating and Refining Study Questions

Purpose:

To ensure that the study questions are relevant to participants' needs, well chosen, well stated and can be empirically answered.

Standard:

Study questions should be clearly defined, articulated and reviewed by a wide audience to ensure that they address the critical aspects of the issues under investigation.

Guidelines:

To meet this standard, the study questions should:

a) Address issues of central importance to participating countries.

b) Be clear in their meaning, implications and assumptions.

c) Reflect knowledge of relevant literature.

d) Be answerable through practical data collection activities.

e) Be capable of further refinement as research planning proceeds.

f) Eliminate bias as fully as possible.

g) Anticipate and respond to unintended outcomes.

h) Break down larger problems into their constituent parts.

i) Be prioritised in order of importance.

Checklist: ☑

Ensure that:

- ☐ Policy-makers, educational practitioners and funding agencies from participating countries as well as international sponsoring organisations have an early opportunity to suggest and review the study questions.
- ☐ Study questions do not make false assumptions.
- ☐ Each study question is really just one question.
- ☐ Each study question does not beg another question that must be resolved before the original question can be addressed.
- ☐ Study questions do not attempt to resolve non-empirical problems by empirical means.
- ☐ Each question can be operationalised in terms of variables that can be validly and reliably measured.
- ☐ Study questions have the same meaning for different persons.

Standard for Designing an IEA Study

Purpose:

To provide guidance in specifying the design of an IEA study.

Standard:

The initial plan for an IEA study should be supplemented by a detailed study design at the earliest possible time. The design should show clearly how the study questions are to be addressed by the study, and provide a plan for the collection, processing and analysis of data, and the reporting of results.

Guidelines:

To meet this standard, a study plan will need to:

a) *State the study questions in sufficient detail* to permit the development of a design to address them.

b) *Prescribe the methodological approach of the study and show how the study will address the study questions*. The kinds of data that will be collected and the inferences that can be made from them should be explained.

c) *Ensure that the sampling plan identifies the target populations* of the study, and that it proposes a sampling design that will provide precise and economical estimates of population parameters. The minimum acceptable levels of sampling precision should be clearly specified. These should be presented in terms of confidence intervals for means or percentages. For example, some IEA studies in the past have specified a 95 per cent confidence interval of ± 5% for sample percentages as a minimum requirement.

d) *Identify which data-collection instruments need to be adapted or developed*, and also outline the scope of this effort and the resources necessary. This work should include plans for fieldtesting instruments as necessary.

e) *Specify the data collection methods to be used*, including the sources of the data to be collected, when and how often data are to be collected, and the collection and processing procedures to be followed.

f) *Clearly outline the data-analysis requirements of the study.* Analysis plans should be provided both for the field-test studies, showing how field-test data will be analysed to inform instrument development, and for the main data collection. The design should specify the analytic methods that will be applied to the data in order to address the study questions.

g) *Include a description of the methods themselves and the software* through which they will be implemented.

Standard for Developing a Sampling Plan

Purpose:

To ensure that the data provided by study sample(s) from each country represent the international target population(s) with a level of accuracy that allows the study questions to be answered.

Standard:

The sampling plan should ensure that national target population(s) are defined operationally in such a way that they can be represented accurately using sample survey methodology. The sampling design should provide a detailed plan for estimating the parameters of the target population(s) with samples of the smallest possible size and greatest possible cost-effectiveness, while maintaining an acceptable level of accuracy.

Guidelines:

To meet this standard, ensure that:

a) *The target population(s) have comparable definitions across countries.* Traditionally, IEA has specified an ideal population known as the INTERNATIONAL DESIRED POPULATION that participating countries were expected to define operationally for their own countries. This national definition is known as the NATIONAL DESIRED POPULATION. For example, the IEA Reading Literacy Study (Elley, 1992, 1994) chose as one of its international desired populations the grade level in school with the greatest proportion of 14-year-olds. In Ireland, the corresponding national desired population was identified as the second year of post-primary school, whereas in the United States the national desired population was ninth grade.

b) *The target population(s) are defined operationally* so as to be amenable to effective probability sampling. In IEA studies the national defined population is the operational definition of the national desired population, and is usually defined in terms of a list of all schools in the country containing eligible students. The national defined population is effectively a sampling frame from which probability samples of schools (and ultimately students) may be sampled.

c) *Any discrepancy between the national desired and defined populations is clearly described*. Such excluded populations may include students in special schools, or in special classes in ordinary schools. These populations should be kept to a minimum. In many countries, students in special schools or classes amount to less than 5 per cent of the age cohort.

d) *A determination is made as to whether it is necessary to obtain data from every member of the population*, or whether sampling techniques

should be employed. In small countries it is sometimes convenient to include all members of the population in a data collection.

e) Where sampling techniques are to be employed, *the required sampling precision or margin of error for the main study is explicitly stated*. This is necessary to enable realistic sample sizes to be planned for each country. It is often helpful to express the margin of error in terms of the size of a simple random sample that would achieve such a margin. This is known as the EFFECTIVE SAMPLE SIZE. For example, a simple random sample of 400 elements would give a 95 per cent confidence interval for a mid-range percentage (i.e., around 50 per cent) of ± 5 per cent. Therefore, the effective sample size for this level of precision is 400. Although real school-based sample designs rarely approach the precision of simple random sampling, the relative precision of a particular design can be quantified by a statistic known as the DESIGN EFFECT, and an estimate of this design effect can be used to help in planning sample sizes.

f) *The minimum sample size for each country is established*, based on the required sampling precision and the nature of the population to be sampled in the country. If the design effect for a country can be estimated (perhaps from a previous, similar study), then this can be used as a multiplier of the effective sample size to find the sample size that would give the required level of precision in that country.

g) *The sampling plan is considered as an integral part of the study design* and not a separate stage to be undertaken after the design has been completed.

h) *Estimates of likely sampling error for key statistics are computed* (or inferred) from previous studies where possible.

i) *The sampling design ensures that the sample size is large enough to represent, with adequate precision, all sub-populations of interest*.

j) If the study has multiple study questions that require different sampling designs for optimal sampling, *a design is developed that represents the best compromise within the range of tolerable sampling errors, costs, burden and other considerations.*

k) *The international co-ordinating centre provides a sampling manual* that itemises each step of the sampling process, and contains detailed tracking forms that should be used by national co-ordinators to document each step of the procedure.

l) Wherever possible, *sampling software is provided to national co-ordinators* to assist them in their sampling activities. Such software not only reduces the burden on national co-ordinators but also produces the tracking information that will be needed for computing sampling weights and for auditing purposes.

Checklist: ☑

- ☐ Describe the defined target population as completely as possible. In IEA studies, the population is usually some cohort of students, defined by age or by grade-level.
- ☐ Define precisely those elements of the population that should be excluded from the sampling process. These may be particular kinds of schools and students with particular disabilities within schools.
- ☐ Specify and justify the sampling technique (e.g., simple random sampling, multi-stage sampling, cluster sampling, probability proportional to size, etc.).
- ☐ In multi-stage samples, define the stages and the units to be selected at each stage. In most countries, the first stage will be schools, but in some very large countries the first stage may be a geographic unit such as a school district.
- ☐ Describe the sampling frame with regard to the source, reference date, number of units, and coverage of the population. The sampling frame is usually a list of schools.
- ☐ Describe stratification and/or clustering techniques, and show their probable effect on sampling precision. Sampling tables showing effective sample sizes for various levels of design effect can help national co-ordinators understand the impact of the sampling design on their survey precision.
- ☐ Justify the proposed sample sizes, and describe the effects of the sample size on the precision of the expected population estimates.
- ☐ Describe the procedures for allocating sample sizes at each stage, with due regard for the types of analyses being planned.
- ☐ Specify the number of sampling units to be selected at each stage.
- ☐ Specify a within-school sampling scheme that is appropriate for the study aims. Random sampling of individual students from all classes across a grade level is an efficient method of sampling students, as it also supports between-school analyses. Sampling intact classes may be more appropriate when the aim is to study classroom processes. National co-ordinators often prefer intact class sampling because it causes less disruption within the school, even though classroom samples typically have larger design effects than individual student samples.
- ☐ Develop procedures for dealing with non-response, and describe the likely impact of non-response on population estimates.
- ☐ If replacement sampling units (e.g., schools) are to be permitted, specify the conditions under which they are acceptable and make explicit the consequences.
- ☐ Specify the minimum sampling participation rates that are acceptable and the consequences of failing to achieve them.
- ☐ Specify how sampling weights will be computed.
- ☐ Specify how estimates of sampling error will be computed.

Standard for Choosing Data Collection Methods

Purpose:

To ensure the selection of the most appropriate and effective methods for collecting data to address the study questions.

Standard:

The data collection methods should be chosen so as to provide the information necessary to address the study questions in a manner that minimises the burden on the data providers (i.e., questionnaire respondents, test takers, etc.) and makes the least demand on available resources.

Guidelines:

To meet this standard, the data collection plan and methods should:

a) *Be determined by the information required to answer the study questions, by what will cause the least burden to respondents, and by available resources*.

b) *Specify the following*: (i) sources of data; (ii) when and how often data are to be collected; (iii) the collection, processing, analysis and reporting procedures to be used; and (iv) the potential descriptive, comparative or causal inferences to be made from the data.

c) *Whenever possible, encompass alternatives to the methods of data collection* that have traditionally been used in IEA studies (namely, group administration of student achievement tests and self-report questionnaires, and questionnaires to teachers, principals and national co-ordinators). For example, it might be feasible to collect information from principals by means of a telephone survey. Advances in video-recording technology have made classroom recordings a more practical proposition, which should make data collection based on direct observation (of pre-recorded video tapes) more affordable.

d) *Specify whether it is necessary to measure change over time or from one study to the next*. Some studies, such as the Second International Mathematics Study (see Burstein, 1992) have incorporated a pre-test/post-test design to permit the study of the effects of classroom instruction. More recently, the Third International Mathematics and Science Study-Repeat is planning to measure trends in eighth-grade mathematics and science achievement from 1995 to 1999.

e) *Include provision for measuring the reliability and validity of the data collected*.

Standard for Developing a Quality Assurance Programme

Purpose:

To ensure that all aspects of the study are conducted to a high standard, and that the data collected are of sufficiently high quality to support authoritative international comparisons. Although quality control should be a feature of all work at both international and national centres, it is particularly important for activities such as test administration, which may be conducted by school personnel and therefore outside the control of study staff.

Standard:

The international co-ordinating centre (ICC) should ensure that all operational documentation emphasises quality control as an integral part of every activity. In particular, data collection activities should include provision for independent monitoring of a sample of data collection sites. Such independent monitoring should be conducted by the national centre and also, as far as funds permit, by the ICC (see also Standard for Implementing Data Collection Quality Control Procedures).

Guidelines:

To meet this standard, it is necessary to:

a) *Develop procedures for checking and rechecking materials,* documents, procedures, analyses and report drafts at all stages of the study.

b) When developing schedules and budgets, *make explicit provision for resources for checking and reviewing all materials*.

c) *Ensure that all materials to be sent from the ICC* to national centres, or from national centres to schools or to the public, *are reviewed by at least one senior member of staff* apart from the original author before dispatch.

d) Given that IEA studies are crucially dependent on accurate translation of survey materials (tests, questionnaires, manuals, etc.) from the study language (usually English) into the vernacular of participating countries, *ensure that the ICC makes every attempt to verify the accuracy of the translations.* This also is necessary to ensure that the survey instruments actually used in participating countries conform to the international standard, and that no bias has been introduced into survey instruments by the translation process.

e) *Ensure that data collection activities are monitored by quality control observers*, who should make unannounced visits to a sample of data-collection sites (usually schools). Since data collection is the responsibility of national centres, they should make provision for such site visits in their plans and budgets. However, the ICC also has a responsibility to ensure that data collection procedures are uniformly followed in all countries, and so should also arrange for independent site visits to a sample of sites.

f) *Ensure that all data analyses are independently checked* and, if possible, recomputed by another analyst using different software. The results of analyses should be reviewed by senior staff members for plausibility and accuracy before being disseminated to a wider audience.

Standard for Preparing an Analysis Plan

Purpose:

To provide a detailed plan that clearly shows how the study data will be analysed to achieve the study objectives. The plan should enable NRCs and other reviewers to evaluate the intended analytic techniques and statistical procedures in the light of the study's aims, scope and resources.

Standard:

The analysis plan should list each of the study's goals or research questions, and provide details of the techniques to be applied to address the issues, together with a justification for each technique. Sufficient detail should be provided to enable reviewers to evaluate the utility and suitability of the application.

Guidelines:

Take into account the following guidelines when developing an analysis plan:

a) *The plan should address, as its first priority, the stated aims of the study*. For each specific research question there should be detailed specifications of how the study data will be analysed to address the issue. If additional analyses are anticipated, these should also be specified, but the priority should be to fulfil the aims of the study.

b) *The specification of analyses is greatly facilitated by the preparation of table shells or dummy tables* that sketch the statistics to be reported in each table of the international reports. These dummy tables help ensure that the analyses concentrate on the statistics necessary to address the main research questions.

c) Generally speaking, *the plan should provide first for simple descriptive analyses* that will produce accessible and easily communicated results, *before proceeding to more powerful analyses*, which often require complex techniques that are more challenging to explain.

d) *The most appropriate analytic or statistical technique should be chosen*, bearing in mind the current state of the art, the quality of the data and the analytic resources available.

e) *The plan should make clear how the analytic techniques chosen will match the design of the study* so as to provide the maximum descriptive and explanatory information. The plan also should make explicit the kinds of inferences that may be made from the data analyses, and the limitations imposed by the design and the analysis.

Developing Data Collection Instruments

Introduction

The emphasis in this section is on standards designed to contribute to the validity and reliability of the study results.

This set of standards applies to the construction of instruments for measuring student achievement in a variety of subject areas, and the development of questionnaires for collecting background information from students, teachers and schools. These standards are designed to ensure that IEA develops tests and questionnaires in which the knowledge, skills, abilities, or personal characteristics measured, the procedures followed and the criteria used will be appropriate to the use for which the instrument is designed. As such, the emphasis is on standards designed to contribute to the validity and reliability of the study results.

More particularly, this section of the Standards covers issues and processes of particular concern to instrument developers, including:

- the framework or domain for the instruments
- the detailed specifications for the instruments
- the development of achievement tests
- the development of questionnaires
- verification of the translations of instruments in the different languages of the participating countries
- field testing of the data collection instruments and procedures.

As noted in the previous section, many of the parameters of instrument development are set by the study design. As part of the planning process, the study design should identify the ages/grades of the target populations, when and how often the data are to be collected, and the test administration and data-processing procedures to be followed. The study design also should identify which data-collection instruments need to be adapted or developed, and outline the scope of this effort and the resources necessary. This process should include plans for field testing instruments as necessary. In essence, the instrument development process begins with a set of basic information about the instruments to be developed, including:

- intended uses of the test
- the curriculum or subject area to be assessed

- the population that will take the test (including anticipated major subgroups)
- the methods that will be used for data collection
- the inferences to be made from test performance.

The initial step in the instrument development process is to define the organisation of the subject matter to be assessed.

The initial step in the instrument development process is to define the organisation of the subject matter to be assessed. This entails modelling or structuring the subject area in a way that can provide the basis for instrument development, scoring and reporting study results. A framework or model of the important aspects of the curricular aims within the subject area (as reflected across the participating countries) generally provides this organisation.

The curriculum frameworks usually include a content dimension consisting of a breakdown of the subject matter into varying levels of specificity and an ability or process dimension that describes the kinds of performance that students will be expected to demonstrate while engaged with the content. Depending on the subject area being assessed, the framework or conceptual model also may need to reflect other curricular goals that focus on the development of students' attitudes, interests, habits and motivations within the subject area.

Once a framework has been developed, there still are many details that need to be decided before the concepts depicted can be transformed into the measurement instruments. These include:

- specifications about the format and response modes to be used
- the timing associated with the instruments
- the weighting associated with the aspects of the framework
- the numbers and types of items
- details about the kinds of items to be included
- guidelines on how to be sensitive to the different cultures represented by the participating countries.

The tests and questionnaires are developed in accordance with the content and statistical specifications. Each cognitive item needs to measure an important topic in the framework and be clear and accurate. The test as a whole needs to provide coverage of the framework and to adhere to overall concerns about difficulty and timing. Questionnaires need to address precise information needs and to contain no more questions than are necessary to obtain that information in a valid and reliable way. Both tests and questionnaires should have sufficient items from each design domain to provide reliable measurement scales. Prior to field testing and for the actual data collection, the instruments need

to be translated into the different languages of testing represented by the participating countries. Those translations, in turn, need to be verified to ensure that the meaning and difficulty level of the items have not changed from the international version.

Field testing serves as a dress rehearsal for the actual data collection effort, giving the participants practice in the sampling, test administration and scoring aspects of the study.

It is important to field test the procedures and data collection instruments before their use in the main study. Field testing serves as a dress rehearsal for the actual data collection effort, giving the participants practice in the sampling, test administration and scoring aspects of the study. It also provides a way to try out the test items to see if they have flaws. Flawed items are either discarded or revised to correct their defects. Another major reason to field test is to obtain reliable item statistics to serve as predictions of how the items will perform when used in the final forms. This information is used in assembling the final test forms, to ensure that they meet both content and statistical specifications.

Finally, subsequent to the test administration, it is important to document the reliability of the tests, and to thoroughly review the item statistics for the final form to detect any potential flaws in the test items or unfair advantage to particular groups of respondents.

Standard for Developing Assessment Frameworks and Conceptual Models

Purpose:

To provide a structure for the study that ensures the instruments developed address the research questions, reflect recent curricular emphases and learning objectives across the participating countries and include what various scholars, practitioners and interested citizens believe should be in the assessment. If necessary, the instruments should also maintain ties to previous assessments to permit the reporting of trends in student achievement, backgrounds and experiences across time.

Standard:

The framework or conceptual model underlying instrument development for the study should define the subject area domain that will be assessed in the study (e.g., mathematics, science, civics, etc.). It should:

(i) Describe the underlying constructs and the procedures followed for defining the domain to be assessed.

(ii) Specify the dimensions, areas, topics and subtopics within the domain that are age/grade appropriate for the target population.

(iii) Explain the purpose for which the results from the tests and questionnaires should be used.

(iv) Provide structure for describing what students should know and be able to do in the subject area being assessed. This structure will provide the basis for developing specifications for developing cognitive items and questionnaires and, eventually, for reporting the study results.

Guidelines:

To meet this standard, ensure that the framework or conceptual model:

a) *Portrays the structure of school curricula in the subject area across the various participating countries.* Measuring student achievement in an international context is what makes IEA studies different from many national surveys, and the variation across countries must be addressed in the framework or conceptual model used.

b) *Reflects experience* gained from previous assessments, research in cognitive development in the areas being assessed, research in test development, and current important developments and documents in relevant fields.

c) *Takes into consideration the status of national reform efforts* across the participating countries in curriculum, instruction and assessment.

d) *Is sufficiently clear* so that knowledgeable experts can judge items in relation to the domains they represent.

e) *Reflects substantive contributions from qualified persons* from the various countries participating in the study. These individuals should represent relevant perspectives, professional specialities, population subgroups, institutions, educational organisations and agencies. This representation often is accomplished by using a committee of such experts to guide framework development. The relevant qualifications and characteristics of each individual involved should be documented.

f) *Is progressive and flexible in viewing the domains to be assessed* so as to include students' ability to connect knowledge and skills between and among the various areas of the discipline. This practice contrasts with that of forcing the subject area into a rigidly structured content-by-ability-level matrix that can distort the nature of the discipline.

g) *Is specific enough* to be useful in the instrument development process, but is not so atomised that the framework or conceptual model obscures the essential themes within the domain being assessed.

h) *Presents a clear description of what is to be tested,* including the critical content to be measured and the relative weight to be given to each part of the domain that is to be measured. The relative weights reflect the importance that content specialists place on the particular kinds of knowledge or skills that the assessment is designed to measure. If the framework covers more than one target population, this guideline needs to be done separately for each population.

i) *Incorporates the comments of reviews by qualified persons* from the participating countries. Such reviews may be supported by procedures such as focus groups of persons from pertinent constituencies, such mail reviews and public hearings.

j) *Is published in an accurate and timely manner*. The framework should be made available to the participating countries and other interested parties before data collection. It is important that the education ministries and agencies funding the study, and the school administrators being asked to give of their teachers' and students' time to participate in the study, have a full understanding of the study goals and what is being assessed.

Standard for Developing Specifications for Tests and Questionnaires

Purpose:

To ensure that the instruments will reflect the framework developed to assess the subject area, are consistent with the resources available for the study and will provide data that can be analysed to address the study's research questions. There also needs to be an assurance that the inferences drawn from those results are likely to be consistent with their intended purpose.

Standard:

The specifications for the tests and questionnaires should describe the content of the instruments, the types of items to be used, and the timing and the conditions (e.g., the physical environment, reference materials, laboratory equipment, etc.) under which the instrument is administered. They should also describe how the responses will be scored.

Guidelines:

It should be emphasised that there is considerable interplay among these guidelines with decisions in one area often influencing specifications in other areas. The goal is to develop valid and reliable instruments within the constraints of the available resources and limits of measurement technology. Thus, decisions about the data collection format, the number and types of items, the timing and the scoring procedures all interact with one another to determine the scope of the instruments possible within the resources available.

To meet this guideline, the item specifications need to include the following information:

a) *A description of the medium and format of the assessment instruments and the response form*. In particular, instrument developers will need to specify: (i) how the directions and test or questionnaire items will be presented to the respondents (e.g., printed test booklets with graphs and tables, simulated interviews on videotape, a series of exercises with feedback on a certain kind of computer); (ii) how and where the respondents will respond (e.g., in separate essay booklets, at work-stations containing science equipment, on audio tape, on stage, or on papers that will be inserted into a portfolio).

b) *An explanation of the design for the instrument development*, including the number of test forms to be developed, the total amount of testing time and the amount of testing time planned for each student.

Similarly, the questionnaires to be developed should be described as well as the time allotted for each. The time requirements should be consistent with the assessment's purpose, but not overburden participants. The age and abilities of respondents should be taken into consideration when estimating response burden.

c) *The kinds of questions or tasks that should be in the tests or questionnaires*. For example, if students are to read materials, the specifications should include:

- information about the lengths of the passages to be read
- whether or not the materials should include pictorial or tabular material
- the reading difficulty of the passages
- the topics of the passages
- whether or not the passages are to be naturally occurring or written specially for the test.

The specifications also need to include information about the tasks students should perform based on their reading of the material, for example, answer questions about particular pieces of information, draw a diagram of a specific phenomena, or make an argument either for or against the issue presented.

Furthermore, the questionnaire specifications should include the topics to be addressed and, if scales are to be constructed, the number and type of items for each scale.

d) *The item types to be used*, including sample items if appropriate. Depending on the format and response form of the assessment, item types can include, but should not be confined to, multiple-choice, constructed responses (both short-answer and longer-answer or problem-solving questions), essay, performance tasks, various types of rating scales, and portfolios.

Multiple-choice items have served IEA studies well, and they are likely to continue to do so. Testing conditions can be easily standardised, the administration costs are comparatively lower than with other item types, and the speed of machine scoring allows very large samples. However, educators are rarely content to assess student performance solely on the basis of multiple-choice questions, and so for a more valid assessment it is usually necessary for other, more complex item types to be included. The mix of test items in the assessment needs to reflect the types of student performances defined in the assessment framework.

e) *An estimation of the number of items or tasks that should be included across all forms of the test*. Increasing the number of tasks permits a better sampling of the domains and generally increases validity and reliability. However, this also affects testing time and scoring costs. Issues of depth versus breadth may enter here. Sometimes it is not possible to adequately cover the framework within the available resources, especially when taking into consideration the types of items to be used and the number of students that needs to be assessed.

f) *A description of any special equipment* that is to be included in the assessments, such as calculators, computers, science kits, manipulatives, books, pamphlets, etc.

g) *The desired psychometric characteristics of the tests and questionnaires*, including:

- the intended level of difficulty of the test
- requirements regarding the target distribution of item difficulties
- requirements regarding the homogeneity of items within test or sub-test forms
- the number of items to be administered to individual test-takers to meet reporting goals
- any other requirements for scaling, equating or measuring trends across time (see also Standard for Analysis Plans).

h) *The procedures and materials that will be used to reflect the differing cultural backgrounds of students* from the participating countries.

i) *A description of the item classification system to be used for documentation purposes*, including aspects of the framework, item type and any other characteristics that will be important in describing the assessment instruments and analysing the assessment results.

j) *The procedures for scoring*, especially when judgmental processes are to be used. The specification must include the approach to be taken (e.g., holistic, analytic, etc.) and the philosophy underlying the scoring. It is sometimes necessary when determining if the scoring criteria are appropriately easy or difficult to adjust the scoring on the basis of pre-test information. For certain criterion-referenced tasks, however, the standard of competency is inherent in the task. Finally, the scoring approach must yield reliable results. Thus, it should include specific instructions for scorers and clear examples of responses to define the score categories.

Standard for Test Development

Purpose:

To ensure that the tests developed by IEA provide fair and accurate measures of students' achievement on the subject matter domain defined by the framework and that they adhere to the test specifications.

Standard:

The tests as a whole should meet domain definitions and test specifications. All aspects of the test need to be clear and accurate, including the directions, stimulus materials and items. The individual items and tasks must be appropriate to the purpose of the test, the population of respondents and the specifications for the test. The procedures used to develop the test must be clearly documented.

Guidelines:

To meet this standard, it is necessary to ensure that:

a) *The directions for each test as a whole and for each item are appropriate and complete*. Respondents should be able to understand readily what they are to do, where they are to respond, and the manner in which they are to respond. If appropriate, include a sample item in the directions that will help the respondents understand what they are to do.

b) *Each item meets appropriate technical standards*. That is:

- The item measures a topic or domain called for by the test specifications.
- Within the study constraints, the item represents the best way to test that topic.
- The item is not confusing or unnecessarily difficult.
- The difficulty level of the item is appropriate for the purpose of the test.
- The item is worded clearly and appropriately for the population, in terms of conciseness, reading level, grammatical correctness and cultural sensitivity. Negatives are avoided if possible.
- The intended answer is accurate and clearly defined, and there are no intentional clues about the answer in the stem or distractors (for multiple-choice questions).
- The stimulus material is clear, correct and can be reproduced in the test. The time taken to read the stimulus material is justified by the items on the test.
- Any inter-dependence from one item to another in a set is appropriate.

c) *Subject matter and test development specialists* who are familiar with the specifications and purpose of the test and with its intended population *have reviewed the test items and scoring procedures for accuracy, content appropriateness, suitability of language, and difficulty*.

d) *The scoring criteria* established for each item *are reasonable and can be reliably implemented* within the parameters of the study.

e) *Test editors have reviewed the items to ensure that the content and phrasing of each item are clear and appropriate*, and that the typography, format, layout and response method do not hinder the task of the test takers.

f) Representatives of the countries and populations participating in the study have reviewed the items so that *language, symbols, words, phrases and content that are generally regarded as sexist, racist, negative toward cultural groups, or otherwise potentially offensive, are eliminated*.

g) *Each test as a whole represents an adequate and appropriate sampling of the domain of knowledge and abilities to be measured*. In addition, the test should adhere to the detailed set of content specifications and meet statistical specifications, such as for mean difficulty level.

h) *The time requirements are consistent with each test's purpose* so that time is not a decisive factor in performance for the large majority of test takers, except for tests designed to measure rate of performance.

i) *The overlap among items is minimal*. Overlap among items in a test has several disadvantages, including the reduction in the range of material sampled and the fact that one item may be answerable on the basis of information contained in another item. The latter is of particular concern when reviewing the items contained in each form of the test.

j) When feasible, *appropriately modified forms of tests or administration procedures are made* so they are available for test takers with handicapping conditions.

k) *The procedures for developing the test are documented*.

l) *The classifications of each item are documented* in accordance with the item specifications and the domain that the item represents.

m) *Permissions have been obtained for any copyrighted materials* that are to be included in the test.

n) *Each test has received a final review and proof-reading* before being sent to the participating countries for translation into the languages of testing (see also Standard on Translation and Translation

Verification). This final review should be performed by someone other than the person responsible for assembling the test.

o) If the test contains constructed-response items, *a plan and materials on how to score these items reliably have been prepared for training representatives* from participating countries.

p) For items that require scoring, *a plan has been developed for countries to obtain information about the reliability of their scoring procedures*. Subsequent to the test administration and scoring, it will be necessary to compute measures of the degree of the reliability of scoring in each country and to review these statistics to determine if there are any flaws in items or scoring procedures.

q) Before analysing the test results, *a thorough item analysis has been conducted to check that the test length is appropriate for the scheduled time available for testing*. The item performance also should be checked country-by-country for appropriate levels of difficulty and discrimination, for any potential flaws in presentation or translation, and for differential functioning among subgroups of respondents by country and gender.

   Flawed items should be removed from the test (either for all countries or for the individual country with the problem) before analysing the results. Similarly, response distributions for some questionnaire items may indicate that they are defective (e.g., confusing, lacking in respondent variation, inconsistent with other items or difficult for the respondents to answer because of confusing presentation).

r) Before reporting, *there is an assurance that test scores, including sub-scores and combinations of scores, are sufficiently reliable for their intended use*. Here, it will be necessary to provide information on reliabilities, standard errors of measurement, or other equivalent information (e.g., information on classification consistency), so that test users also can judge whether reported test scores are sufficiently reliable for their intended uses.

Standard for Questionnaire Development

Purpose:

To ensure that questionnaires developed for IEA studies address the issues specified in the study questions in a way that maximises the reliability and validity of the measures while minimising the burden on respondents.

Standard:

Questionnaires should be clear, simple, concise and manageable.

Guidelines:

The following are guidelines to ensure the quality of questionnaires and to maximise response rates:

a) *Be specific and think in terms of the results to be reported*. The goal is to define precisely the information desired and to write as few questions as possible to obtain it. In the final questionnaires, avoid questions peripheral to the study aims (see also Standard for Pre-testing Data Collection Instruments and Procedures).

b) *Think whether the questions will produce credible information*. Are respondents able to answer the question? Will they answer the question?

c) *Review each question carefully to ensure that it relates to one idea only*. If there is more than one idea, use more than one question. Also, check if there a simpler or more direct way to ask the question. Revise or delete any confusing words, including those that may be unfamiliar to respondents, have more than one meaning or are especially difficult to translate into other languages.

d) *Avoid open-ended questions in the final questionnaires*. Open-ended questionnaire items increase the scoring costs associated with the study, increase respondent burden and suppress responses from the less literate segments of the population or from respondents who are less concerned about the topic at hand.

e) *For questions with a range of response categories*, make sure that the categories are mutually exclusive and that there is a response that applies to each of the respondents. *For questions that do not apply to all of the respondents,* use directions to tell people to skip to a later question. (With questionnaires for upper secondary students and adult respondents only, avoid these filter questions for respondents under 16 years of age.) Also, keep the number of response categories or scale categories to a reasonable number. The distinction among categories must be useful to the study and easily distinguished by the respondents.

f) When a number of questions all use the same response categories, *ensure that the questions are appropriate to the responses*. Also, avoid placing together questions with similar, but not identical, response categories or options.

g) *Design the layout of the questionnaire so that it is easy to follow and the response options for each question are clear*. Cramming as many questions as possible into the fewest number of pages is a poor strategy for encouraging high response rates. Lengthy questionnaires similarly are not likely to encourage high response rates.

h) *Arrange the questions so that the flow from question to question is natural and sensible*, and the skip patterns are clear. Some variety in the response formats can improve interest and attention.

i) *Consider whether matrix sampling may be appropriate in the development of the questionnaire*. Matrix sampling in this case means that not all respondents are asked all questions. Although this method will add some cost and complexity, it can greatly reduce response burden. Matrix sampling should only be considered if the study objectives can be met with adequate precision.

j) *Use small-scale pilot studies* to try out question wording, and ensure that items intended to form a scale have appropriate psychometric characteristics for reliable measures.

Standard for Translations and for Verifying Translations

Purpose:

To ensure that (i) the cognitive items are translated from the international versions into the target languages without changes in meaning or difficulty; (ii) that cultural differences are kept to a minimum; and (iii) that the meaning and content of the questionnaire items are retained through translation. The goal is to obtain translated instruments of high quality that will provide comparable data across countries and cultures.

Standard:

When translating test items or modifying them for cultural adaptation, the following must remain the same as the international version: the meaning of the question; the reading level of the text; the difficulty of the item; and the likelihood of another possible correct answer for the test item.

Guidelines:

To meet this standard, ensure that:

a) *Participants use a system for translating the tests and questionnaires that incorporates an independent check on the translation*. One such system might be that, for each subject matter, each national centre engages a minimum of two translators who are subject matter specialists with fluency in the source language and the target language. The specialists produce two independent translations of the items, and then have a third party compare the two versions. When there are differences between the two versions, the best version of the translation is selected.

In general, the translators' work includes the following:

- Identifying and minimising cultural differences.
- Finding equivalent words and phrases.
- Making sure that the reading level is the same in the target language as in the international version.
- Making sure that the essential meaning of the items does not change.
- Being aware of changes in layout due to translation.

b) *The international co-ordinating centre (ICC) explains the difference between appropriate and inappropriate cultural adaptations.* In general, cultural adaptations should be confined to units of measure, proper nouns, common nouns, spelling, verbs (not related to content), usage and punctuation.

c) *The survey operations documentation includes materials* for the participating countries *that describe the translation and cultural adaptation procedures and the process for translation verification*.

d) *Countries record any deviations in vocabulary, meaning or item layout from the international versions and forward them to the ICC*. These can be sent as the countries are translating and adapting the items, which permits the national centres to receive approval of adaptations quickly and expedites the preparation of the test booklets and questionnaires. Make sure that a subject matter specialist approves all such deviations.

e) *The ICC, acting independently of the countries, has the translations and adaptations of the tests and questionnaires verified by professional translators*. For each country, a professional translator should review the overall layout of the instruments, the translation of the student instructions and the translation of each item. The professional translator should compare each translated item with the international version and document the differences from the international version.

f) *The national centre addresses the deviations that the translation verifier judges as affecting the results.* Examples include incorrect ordering of response options in a multiple-choice item, mislabelling of a graph that is essential to a solution, or an incorrect translation of a test question that renders it no longer answerable or indicates the answer to the question.

g) *The ICC verifies that all deviations judged to affect the results have been rectified*.

h) Subsequent to the testing, *the ICC conducts a thorough review of the item statistics* to detect any errors in the translation or adaptation verification process that were not corrected. This process also should be carried out following field-testing, so as to minimise the possibility of translation errors in final versions of the test instruments.

Standard for Field Testing Data Collection Instruments and Procedures

Purpose:

To obtain information about the sampling procedures, the test administration procedures, the performance of the test questions and the scoring procedures before the questions are included in the final forms and the procedures are used in the actual assessment. Note: the purpose is to field test[2] the items and procedures, not the students.

Standard:

All achievement and questionnaire items, tasks and directions should be field tested on a population as similar as possible to the target population. The sampling, test administration and scoring procedures to be field tested should be as close as possible to those that are planned for the actual data collection phase of the study. Smaller scale field testing or piloting approaches should be considered before full-scale field testing to ensure that the full-scale field test provides the most accurate information possible.

Guidelines:

For a field test to yield information that can improve the main survey administration, it will be necessary to:

a) *Administer the assessment instruments to a sample from a population in each participating country that is similar to the national target population*. However, the field-test sample cannot include students who will be assessed in the main survey. Because the field-test results are usually not useful to the individuals and/or schools that participate in the field testing, locating the field-test population may require considerable effort. Participants will need to understand that the quality of the main survey data depends on effective field testing, that participation in field tests is necessary, and that the quality of the sample used for field testing is extremely important.

b) *Field test a sufficient number of items to permit assembly of the final forms from the field-tested items*. This number can encompass from 150 to 300 per cent of the required number of items, depending on previous experience, the nature of the content categories and the nature of the item types. When the potential differences among the participating countries are taken into consideration, a good rule of thumb is that about twice as many items as will be needed should be field tested.

2 In this standard, any type of experimental tryout or trial of new materials before the administration of the final form is referred to as field testing. Field testing can take many forms from 'think alouds' with 10 or so students to better understand new item types, through pilot testing with a small group of countries, to full-scale field testing that simulates the procedures planned for the final assessment as closely as possible.

c) *Consider using successive small-scale field tests to refine questions and their wording when developing questionnaires*. These also can be used to help refine the process of building attitude and opinion scales.

d) *Consider conducting 'think alouds' with small groups of students when new item types are being used for either the tests or questionnaires*. This will help the test and questionnaire developers refine the items before field testing.

e) *Consider conducting a small-scale field-testing phase before a full-scale field test in situations where a large number of items are needed, experience is lacking and/or new item types are being used.* In the context of IEA studies, this small-scale field test could occur with a subset of the participating countries.

f) *Ensure that timing for the field-test forms corresponds to the timing planned for the final assessment*. This process will provide better information about the length of the tests and questionnaires relative to the time assigned to complete them. However, if the items are for an unfamiliar population or are of an unfamiliar type, try to be generous with the time allowed for each item. A test that requires students to work quickly to complete it often results in little or no data for the items near the end of each pre-test instrument.

g) *Construct (where possible) the field-test versions of instruments so that they can be transformed into final forms with a limited amount of revision*. One way to do this is to prepare parallel sections or blocks of test items. The blocks needing the fewest revisions are selected for conversion to the needed final forms. Another way is to have 'parent' blocks that correspond to those needed for the final form and some additional blocks of replacement items for those areas more difficult to measure.

h) *Ensure that the forms for field testing are reviewed by content experts, test development experts, editorial specialists and proof readers* (see also Standard for Test Development).

i) *Conduct training sessions for field testing that simulate those planned for the actual data collection effort* (see also Standard for Selecting and Training Data Collection Staff).

j) *Emphasise the need for test security* and the procedures to be followed to maintain it during the field-testing process.

k) *Score the field-test responses as if they were from an actual administration of the test*. Doing this will help determine if respondents understand what to do, whether the tasks elicit the desired kinds of response, and whether the responses can be easily and reliably scored with existing criteria and/or scales.

l) *Conduct an item analysis on the basis of the field-test responses*. This analysis should include estimates of item difficulty and item discrimination, statistics about differential item performance by country and by gender, and data about the reliability of the scoring procedures. This information will be used to detect flawed or biased items so that they can be corrected or discarded, to ensure that some respondents do not have an unfair advantage, and to ensure that the final forms meet the statistical specifications for the test.

m) *Check the performance of questionnaire items*. The response distributions for some questions may indicate that they are defective or that there is not enough variation in the population to make the question worthwhile. Also, pairs or sequences of questions may yield inconsistent responses, indicating the need for rewording or changing the response mode. Low response rates may indicate 'speededness' or lack of interest by the respondents. If sets of items are intended to form scales, examine the relationship between these items, and assess the reliability of the resulting scale.

n) *Integrate the selection of the field-test sample into the main survey sampling design and the sample chosen, using the same probability-sampling methods as in the main survey.* Although field tests often are conducted on small convenience samples of schools or classes, it is much better if the field-test sampling is integral to the main survey sampling plan. For example, the Third International Mathematics and Science Study-Repeat (TIMSS-R) selected the field-test school sample (25 schools) from the same sampling frame and at the same time as the main survey sample (150 schools) in most countries.

o) *Remember that school sample sizes for the field test can be much smaller than for the final study, and that the procedure for selecting schools can be less rigorous.* Within schools, however, follow the main survey procedures as closely as possible. This will ensure that the test-item characteristics, the questionnaire-item characteristics and the operations procedures are realistically evaluated in all participating countries.

p) *Ensure that the field-test samples are large enough to provide useful diagnostic statistics for item selection and refinement*. As an example, the TIMSS-R field test required at least 200 student observations for each test and questionnaire item. Given that each student was assigned one of five test booklets, the field test required a sample of at least 1,000 students.

Data Collection and Processing

This set of standards applies to the collection of survey data by means of achievement tests and questionnaires and the processing of that data.

In designing the systems, procedures and materials to be used to collect the data, it is very important to keep in mind that IEA studies rely extensively on school systems and their students and staff to be the primary providers of information. School systems and staff must balance competing demands for their time, attention and resources. Similarly, student assessments interrupt school schedules and limit the amount of time that can be used for instructional purposes. The burden imposed by a study on the proposed participants must be carefully evaluated, with due consideration given to the fact that providing data for the study is generally a voluntary activity and not a primary responsibility for the participants.

The design of data collection systems and procedures for IEA studies must reflect international requirements for uniformity and rigor while recognising that there are national variations in how school systems are structured and operate.

The design of data collection systems and procedures for IEA studies must reflect international requirements for uniformity and rigor while recognising that there are national variations in how school systems are structured and operate. The data collection design must be made operational through carefully defined procedures.

Procedures must be implemented in a standardised manner and documented using survey forms.

- The study procedures and forms must be explained in manuals and other study materials.
- Study staff must be trained to carry out the procedures and use the forms correctly as described in the manuals and materials.
- The procedures, forms, materials, manuals and training programs must strike a balance between national differences and international concerns for uniformity.

Determining when to allow variation among countries and when to require uniformity is a major task in developing data collection procedures and materials. When national adaptations are permitted, the proposed adaptations should be reviewed by the international co-ordinating centre (ICC) for international comparability and for conformity with the study intent and design.

In a similar way, procedures for processing the data must be developed, documented and transmitted by international staff to national staff to make sure that the data from each country are prepared in the same way as data from other countries. Procedures for reviewing and checking source documents also should be specified.

The following sections define standards for each of the major components of data collection and processing. As with the other sets of standards in this manual, the standards generally apply to the international level. However, given that data collection is the responsibility of staff of the participating nations, we have defined standards at the national level as well where appropriate.

Standard for Drawing a Sample

Purpose:

To ensure that the samples drawn in each country comply with the study sampling plan and are representative of the specified target population with minimal risk of bias, and that the sampling procedure will accommodate the computation of sampling weights and variances.

Standard:

Samples for IEA studies must conform to the study sampling design and use methods derived from sound and defensible sampling theory. They must meet specified quality standards in terms of coverage, participation rates and data reliability.

Guidelines:

When implementing the sampling design in a country, sampling consultants and national research co-ordinators need to:

a) *Ensure that the method chosen to implement the sampling design is easy to apply*. Because IEA studies usually involve countries with varying levels of sampling expertise and capabilities, the design is typically one that can be implemented in a straightforward manner. However, procedural simplicity should not be accomplished at the expense of analytical objectives and data reliability.

b) *Ensure that the sampling method used is verifiable*. It should be possible, with readily available documentation, to replicate centrally the sampling that took place in the national centres.

c) *Use sampling forms to document the sampling process*. (The forms are designed to track the sampling process at every step.) This documentation will provide:

- a written record of the sampling process
- a detailed description of the sample design parameters
- descriptive information (such as the extent of coverage, exclusion rates and participation rates) essential in determining the quality of the resulting samples
- the data from which to derive the sampling weights for the computation of all survey statistics.

d) *Remember that for very large countries like the United States or the Russian Federation it may be necessary to use geographical areas such as regions as primary sampling units, and to sample schools as a second stage*. Target populations in IEA studies are usually hierarchical in nature: students within grade levels within schools; or students within classrooms within grade levels within schools. Multi-stage

sampling designs are suited well to this situation, and are used extensively in IEA studies. Often, the nationally defined population (see also Standard for Developing a Sampling Plan) is defined in terms of a list of schools containing eligible students. This list serves as the sampling frame for the first stage of the sampling design, with schools as the primary sampling units.

e) *Avoid the following problems when constructing the sampling frames*:

- UNDER-COVERAGE: This occurs when there are eligible units of the population (e.g., schools, classes or students) not present on the sampling frame. Under-coverage leads to under-reporting of survey results because these results do not apply to the whole of the desired target population. It is important to document the nature and extent of under-coverage.
- OVER-COVERAGE: Over-coverage occurs when the sampling frame contains sampling units that do not belong to the study's target population (e.g., schools that do not contain any of the target grades for the study). Over-coverage can become a serious problem if it leads to the selection of a significant number of sampling units that will ultimately have to be discarded. The result will be a much smaller sample size than expected.
- DUPLICATION: Duplication occurs when sampling units appear more than once on the sampling frame. When this problem occurs, it becomes difficult, if not impossible, to properly compute selection probabilities. Also, sampling a unit more than once can lead to operational difficulties.
- ERRONEOUS INFORMATION: A sampling frame can contain either erroneous information, or simply information that is out of date. Such errors can lead to coverage problems, errors in stratification and inappropriate selection probabilities.

f) *Consider using* STRATIFIED SAMPLING when it is necessary (i) to ensure that all strata are represented in the sample with fixed sampling fractions, (ii) to ensure that adequate sample sizes are achieved in each stratum, or (iii) to minimise sampling error by eliminating between-stratum variance.

g) *Consider using* EXPLICIT STRATIFICATION (where a separate sample is drawn within each stratum) whenever analytical objectives or sample design issues require its use. Different sample designs, using, for example, disproportionate sampling allocations, can then be applied to each explicit stratum. This could be necessary if it were desirable to over-sample small sub-populations relative to other, larger ones.

h) *Note that some of the gains in proportional representation of sub-populations and sampling precision that accrue from explicit stratification can be achieved using a simpler technique known as* IMPLICIT STRATIFICATION. In this approach, the units in the sampling frame are listed together by stratum, and within stratum are sorted by any available variables that are expected to be related to the outcome measure (which is usually student achievement). Sampling is done across the entire sampling frame rather than separately within each stratum, using a random-start, fixed-interval form of systematic random sampling. Implicit stratification can be used if the sole objective is to ensure proportional representation of the implicit strata in the sample. Implicit stratification can also lead to smaller sampling errors if the variables used in stratification are correlated with the characteristics being measured.

i) *Ensure that,* regardless of the stratification methods used, *strata are constructed so as to be large enough to contain several sampled units.* Two sampled units per stratum are the fewest that can be used and still yield within-stratum sampling errors. However, much larger numbers are preferable.

j) *Remember that if the target population definition is strictly age-based, then students should be sampled directly within sampled schools, without reference to their classroom*. In this case, the school is usually sampled with probability proportional to size, then a fixed number of students is sampled within the school. This practice ensures that each student in the stratum will have an equal probability of selection.

k) *Consider sampling intact classrooms if the target population definition is grade-based and the study has a classroom-related focus*. In general, intact classrooms should be sampled with equal probabilities within sampled schools.

l) *Note that if intact classrooms are the preferred sampling unit, it may be possible to sub-sample students within sampled classrooms.* This practice may be appropriate for school systems with very large classrooms where it is desirable to reduce coding and data entry. If student sub-sampling is planned, then classrooms should be sampled with probabilities proportional to their sizes.

m) *Remember that if replacement schools are to be used in place of non-participating schools, they must be identified when the school sample is drawn*. Replacement schools should be identified in a strictly deterministic way, with specific criteria in mind, and each one should be linked to one, and only one, sampled school.

Standard for Planning for Data Collection

Purpose:

To ensure that comparable data are collected in participating countries while recognising national variations in school system characteristics and data availability.

Standard:

Planning for data collection involves implementing the international study design in the form of standardised procedures, forms and materials to ensure comparability of the data collected. All aspects of data collection should be specified in detail in the plan, including selecting respondents and securing co-operation; hiring and training study staff; documenting survey procedures, questionnaires, and forms; and specifying quality assurance procedures.

Guidelines:

In meeting this standard the international co-ordinating centre and the national co-ordinators of the participating countries each need to follow specific sets of guidelines. Note that participating countries have a major responsibility for planning the data collection.

The INTERNATIONAL CO-ORDINATING CENTRE should:

a) *Involve representatives of the participating countries* in specifying the framework for the data collection plan and schedule of data collection.

b) *Develop a detailed data collection plan* that includes the following:
   - Descriptions of the target populations and rules for including/excluding population members.
   - Procedures for selecting school and student samples and identifying other respondents such as teachers and school principals.
   - Instructions for completing all data collection forms and instruments.
   - The administrative steps involved in each data collection activity.
   - Descriptions of allowable variations in forms and procedures and the process for obtaining approval of these variations.
   - Requirements for national monitoring of data collection activities.
   - A schedule of international data collection milestones.
   - Requirements for data confidentiality and security.

- Response rate requirements and procedures for using replacement respondents.
- Requirements for data collection staff organisation and training.

c) *Use the plan to develop manuals* specifying the procedures and materials appropriate for each level of data collection (i.e., national co-ordinator, school co-ordinator, test administrator, etc.). Here, it will be necessary to:

- Give representatives of the participating countries multiple opportunities to review draft plans, manuals, materials and forms.
- Conduct field test(s) of procedures, forms and materials, and make appropriate modifications.
- Develop a training program for national study staff.
- Develop a program for monitoring the quality of the data collection.

The NATIONAL CO-ORDINATORS should:

a) *Carefully review all study procedures and materials* for compatibility with national school system characteristics.

b) *Follow specified procedures for obtaining necessary modifications* in procedures and materials.

c) *Adapt international materials*, as necessary, and get international approval.

d) *Prepare all administrative manuals, materials and forms.*

e) *Develop informational materials and a plan for securing permission and support* from the requisite administrative levels.

f) *Notify participants as early as possible.*

g) *Develop a plan for selecting, hiring and training data collection staff.*

h) *Develop a plan for administration and monitoring of data collection activities.*

Standard for Selecting and Training Data Collection Staff

Purpose:

To make sure that data collection staff have the knowledge and skills to carry out all data collection activities.

Standard:

The qualifications required of data collection staff should reflect the complexity of the tasks they will perform. Training programs in data collection procedures should be sufficiently thorough that trainees complete their training with all the skills necessary to carry out their responsibilities successfully. Training staff should be able to evaluate the progress of each trainee as they work through the training program.

Guidelines:

To meet this standard, the international co-ordinating centre and the national co-ordinators will need to carry out various activities.

Staff from the INTERNATIONAL CO-ORDINATING CENTRE will need to do the following:

a) *Determine whether the study design requires any special training or skills.* If so, the ICC staff will need to develop a training program that includes these skills and then ensure that national staff receive this training.

b) *In situations where special training or skills are required that the ICC cannot provide, make sure that countries are so informed as early in the process as possible*.

c) *Although a particular staffing pattern may not be required internationally, make sure that each country receives the necessary skills and training.* (Some countries will choose to use school staff to collect data and administer assessments, other countries will use national staff, and others will hire and train data collectors specifically for the study.)

d) *Develop manuals for each level of data collection staff*, and make sure that these are inclusive of all procedures and materials and are clearly written so that the staff using them will be able to do so with minimal training.

e) *Centralise some functions*, for example, training of trainers and training of quality assurance staff.

To play their part in meeting this standard, the NATIONAL CO-ORDINATORS should:

a) *Follow the international guidelines*, requesting only minimal modifications to reflect significant national differences.

b) Depending on the level and experience of data collection staff, *consider expanding the international manuals, materials and training plans to cover points not covered or experience that may be lacking.*

c) *Provide a contact person who is knowledgeable about the study* and will be available to answer questions from collectors and respondents during the data collection period.

d) *Ensure that data collectors have the opportunity to practice with the data collection instruments* (e.g., administer the assessment, fill out the forms, transcribe the records before they have to use them). Preferably, this should occur at training.

e) *Hire and train sufficient backup or substitute staff.* If relying on local school staff as data collectors, have alternates in mind.

f) *Debrief staff after field tests and revise nationally developed materials* to reflect the field-test experience.

g) *Monitor field activities and identify problems encountered* so that corrective action can be taken quickly.

Standard for Minimising Response Burden and Non-response

Purpose:

To ensure that data collection activities are minimally intrusive and burdensome and that the data collected represent the population being studied.

Standard:

Since burden and non-response are related to each other, and data collection procedures and materials are based on the study design, all aspects of the study design should be carefully reviewed against realistic expectations of respondents, recognising that participation is generally voluntary and not a primary activity for respondents.

As materials and procedures are developed to implement the design, they should be reviewed for clarity, ease of use and simplicity. Acceptable levels of response and the role of replacements should be defined before beginning the data collection. Standards for reporting data should also be defined before the beginning of data collection.

Guidelines:

The following are guidelines for minimising burden and non-response:

a) *Set realistic limits* regarding acceptable respondent burden for questionnaires and assessment instruments, keeping in mind who the respondents are.

b) *Use the results of field tests* to check that estimates of burden are realistic and acceptable.

c) *Simplify procedures as much as possible* and describe them clearly and carefully.

d) *Make sure that all forms and instructions for respondents are clear* and easy to follow.

e) *Fully inform respondents about the purposes of the study* and how their data will be used.

f) *Request from respondents only information that they alone can provide.* If information can be obtained by study staff from other sources or records, consider obtaining the information from these sources.

g) *Schedule the data collection at the convenience of the respondents*, to the extent possible, recognising that schools are very busy places and that participation in the study is often voluntary.

h) *Consider what the incentives are for participants and feature these incentives in recruiting materials.* Public duty, the opportunity to participate in an important international study, and access to study results or reports can be effective incentives at minimal cost.

i) *Develop procedures for handling non-response and refusals* and train field staff in these procedures.

j) *Make sure that the timeline for recruiting participants allows time to notify and recruit replacements,* if appropriate.

Standard for Implementing Data Collection Quality Control Procedures

Purpose:

To make sure that data are collected according to study requirements.

Standard:

Quality control should be an integral part of the study at both the national and international levels. Quality control encompasses both internal mechanisms that are built into each stage of data collection to ensure that procedures are implemented correctly, and external reviews administered by staff members who are separate from the staff being evaluated.

Guidelines:

The following guidelines apply to both the international co-ordinating (ICC) centre and the national centre:

a) *Incorporate quality control into study procedures, forms and materials* as they are being developed. For example, information can be required from two different sources and compared for consistency. Range checks, valid response checks and logic or consistency edits can be built into forms as well as into data processing.

b) *Include quality control in training programs* by having trainees complete samples of the work they will be doing and submitting the work for review.

c) *Conduct a quality control review of random subsets of data collection forms* to make sure that they have been filled out completely and accurately.

d) *Hire and train quality control monitors* to be responsible for external quality control activities. Ideally, this work should be done independently by both the ICC and the national centre. The monitors should observe data collection activities in a subset of sites. For example, in the Third International Mathematics and Science Study-Repeat (TIMSS-R), the ICC conducted site visits to 10 per cent of sampled schools in each participating country, and the national centre conducted visits to a further 10 per cent. If the resources available will not support actual observations, then in-person or telephone interviews with data collectors, test administrators or other respondents should be conducted to the extent possible. All observations and/or interviews should be conducted according to a defined protocol, and responses documented in a standard form.

e) *Use reports summarising the results of each stage of data collection* to monitor the status of data collection activities.

f) *Take prompt action to identify and correct the causes of quality problems.*

g) *Have each national centre provide a report to the ICC* documenting the extent and results of quality control activities carried out by the national centre.

Standard for Documenting National Data Collection

Purpose:

To ensure that the national data collection activities can be reviewed by the international co-ordinating centre (ICC), and documented for use by later users of the data.

Standard:

The documentation of the data collection for each country should include all adaptations to and deviations from the international procedures that have been made by the country.

Guidelines:

To meet this standard, the documentation should include:

a) *All local modifications made to manuals* distributed by the ICC.

b) *A report on all data collection activities* written by the national research co-ordinator.

c) *The results of the checks* that have been made by the ICC and the international data processing centre to validate the correctness of the data collection.

d) *A report on the quality control site visits* conducted by the national centre.

Checklist: ☑

The ICC should ensure that:

- ☐ Each country has documented its national adaptations in a standard, pre-defined way.
- ☐ The data collection report from the national co-ordinator includes:
  - the national population definitions
  - the sample design as implemented in the country
  - the definition of disability used to exclude students with disabilities
  - all national adaptations to the international instruments and procedures
  - the names of the test administrators and how they were trained
  - how and when the tests were administered
  - details of staff training programs for scoring of free-response items and data entry
  - procedures for monitoring the accuracy and reliability of free-response scoring
  - quality control procedures for data entry (e.g., double coding, data reviews, check programs, rescanning).
- ☐ Each country provides a quality control report on the data collection.
- ☐ Each country provides a set of its data collection instruments.

Standard for Planning for Data Preparation and Processing

Purpose:

To ensure that all participants have the tools and information that they need for data preparation and processing so that they (i) are ready to produce data files for the international co-ordinating centre (ICC) that conform to the international data structure, (ii) are comparable internationally, and (iii) are free from error as far as possible.

Standard:

National co-ordinators should plan their data preparation activities in good time, and ensure that they have all of the information and understand the procedures necessary to create data files that conform to the international data structures. The ICC should ensure that participants understand the importance of planning for effective data preparation, and should provide them with all the necessary tools, information and training in data preparation. All prescribed data preparation procedures and quality control checks should be clearly documented. Training in the use of software and in the application of data preparation procedures should be provided, as necessary.

Guidelines:

To meet this standard, NATIONAL CO-ORDINATORS should:

a) *Ensure that they fully understand all data preparation and processing procedures*.

b) Are familiar with any software that has been provided.

The INTERNATIONAL CO-ORDINATING CENTRE should:

a) *Provide participants with manuals* that give detailed information on:
   - how to enter data (and if necessary provide programs for entering the data)
   - how to assign identification numbers to students, teachers and schools
   - the variables (including valid codes and codes for missing or inapplicable responses) that should be used
   - the data formats and file structures that should be used.

b) *Ensure that participants are familiar with international procedures and data structures*.

c) *Ensure that participants understand the importance of tracking information* for computing sampling weights, cross-checking data and verifying data-collection procedures.

d) *Provide for training and assistance to staff working in the participating countries*. If resources permit, it may be desirable that members of the ICC visit the national centre to provide support and training.

Checklist: ☑

National co-ordinators should:

- ☐ Install all software on the actual machines to be used for data preparation and test them well before entering real data.
- ☐ Recruit data preparation staff and train them in all aspects of their task.
- ☐ Design quality control procedures that ensure all data files are of the highest quality and are as accurate as possible.
- ☐ Ensure that the structure of the data files matches the layout of the survey instruments.

The ICC should ensure that:

- ☐ All participating countries have the tools and the training necessary to prepare the data files.
- ☐ The proposed record identification system for each country meets the requirements of the study and fits with the sampling plan for the country so that all data files from each respondent group (students, teachers, principals, etc.) can be merged together.
- ☐ The variable definitions (and especially their valid ranges) given in the international code book are usable in each country.
- ☐ Communication structures are established that guarantee the national centres will get support from the ICC as needed.
- ☐ Adequate training in data preparation tasks is provided to participants as necessary.
- ☐ Enough redundancies are built into the data to make data quality checks possible.

Standard for Processing and Checking Data

Purpose:

To ensure that data provided by national centres are in a common format and are suitable for international analyses.

Standard:

The data provided by national centres should conform to the international data structure and to the data definitions provided in the international code book. It should also be free of internal inconsistencies. The international co-ordinating centre (ICC) should ensure that all data files provided by participating countries comply with international formats and are free from error.

Guidelines:

To meet this standard, NATIONAL CO-ORDINATORS should:

a) *Ensure that all prescribed procedures for data preparation are closely followed* by their data preparation staff, and all prescribed checks for data integrity and quality control are applied. Any errors or inconsistencies that are identified should be resolved before sending data to the international co-ordinating centre (ICC). It is the responsibility of the national co-ordinator to provide error-free data to the international centre. Data files that do not conform to the international standard may not be accepted by the ICC and may be returned to the originating centre for revision.

The INTERNATIONAL CO-ORDINATING CENTRE should:

a) *Encourage national co-ordinators to check their data carefully as data files are being created*. Errors identified at an early stage are much easier to resolve and do not propagate throughout the data set. Wherever possible, software for conducting consistency checks should be provided to national centres to facilitate early data checking and error resolution.

b) *Check the structure of all data sets provided by participants against the international data structure* and report all deviations to the providers. Problems should be resolved as soon as possible, with the involvement of the national co-ordinator as necessary. In particular, attention should be paid to the following:

- All variables are checked for valid ranges.
- Record identification numbers are checked for uniqueness and integrity. If a hierarchical numbering system is used, all levels of the system should be internally consistent.
- All data files from each country are checked to ensure that all components match correctly.

- All questionnaire data are checked for internal inconsistencies.
- Data files are checked against the national test instruments and questionnaires.
- Item analysis statistics are computed for each cognitive and questionnaire item, and reviewed at the ICC.
- Preliminary univariate statistics are computed for each variable and reviewed for outlying cases and missing values.

c) *Report all extreme or unlikely values for variables* to national centres for investigation.

d) *Send (for review) item analysis and preliminary univariate statistics* to each national centre.

Standard for Documenting Data-Processing Activities

Purpose:

To ensure that all modifications applied by the international co-ordinating centre (ICC) to participants' data are recorded, whether such modifications were the result of structural reorganisation or the result of errors. The integrity of the data should be clearly demonstrated and all deviations from the international standard explicitly documented.

Standard:

National co-ordinators should be notified of all errors and inconsistencies encountered in their data. All modifications to data arising from the resolution of such issues should be endorsed by the national research centre and clearly documented, preferably in a database.

National deviations from international standards should also be clearly documented, to ensure the validity of international comparisons and to facilitate subsequent secondary analysis by researchers who do not have first-hand experience of the data collection.

Guidelines:

To meet this standard, the following need to be fulfilled:

a) *The ICC must ensure that any documentation of data-checking and editing activities* that it sends to the countries for verification include:

- A list of all deviations from the international standard that have been found in the data and instruments.
- A list of all data-cleaning rules that have been applied to the data.
- A list of all edits that have been applied globally to the data files.
- A list of specific edits that have been applied to individual observations.
- A database consisting of the edited data.
- A code book and description for the new data base.
- Computational definitions for any variables that have been re-coded or derived from other variables.
- Item analysis and preliminary univariate statistics for each variable.

b) When all edits and modifications to the data have been finalised, a database for secondary analysis is usually produced. *It is the ICC's responsibility to produce documentation for users of this database* that includes:

- A description of the study design, including achievement test and questionnaire frameworks, instrument design, data collection instruments and the sampling design.
- A list of all data-checking and editing rules applied to the data.
- A code book for each data file giving all the information necessary for a secondary analyst to access and analyse the data.
- An explanation of the structure of the data base.
- A deviation report that lists all adaptations and modifications made by participants to the international instruments or procedures.
- Definitions of all variables that have been re-coded or derived from other variables.

Analysing Data and Reporting Results

Given the cost and the amount of effort required to develop an IEA study and collect the data, it is essential that the data are analysed and reported in such a way that the study aims are achieved and the results widely disseminated.

Given the cost and the amount of effort required to develop an IEA study and collect the data, it is essential that the data are analysed and reported in such a way that the study aims are achieved and the results widely disseminated. The analysis plan should describe in detail the analytic approach to be used and the specific analyses to be conducted. Once the plan has been reviewed and accepted it should be implemented faithfully.

Because IEA studies typically depend heavily on sample survey methodology, it is important that correct sampling weights are computed and applied so that sample statistics accurately represent population parameters. The standard for developing sampling weights provides guidelines for computing these weights and suggests procedures for checking their accuracy. While sampling weights are intended to ensure that sample statistics are an accurate reflection of population values, this estimation always involves some error due to the sampling process. The standard for reporting sampling and non-sampling error makes explicit the need to provide realistic estimates of the magnitude of the sampling error inherent in any sample estimate, and to present these estimates in an accessible and easily understood form. This standard also provides guidelines for reporting sources of bias that may result from non-sampling errors, including deficiencies in population coverage or low participation rates.

IEA studies typically collect data in the form of responses from students, teachers or principals to questions on achievement tests or questionnaires. To facilitate analysis and reporting, these responses are frequently summarised in the form of measurement scales, which are expected to provide more reliable and valid information about the topics under study than would the responses to individual questions. For example, a student's responses to a set of questions in mathematics may be combined to give an overall mathematics score (which summarises the student's mathematics achievement in a convenient and reliable manner), or a teacher's responses to questions about instructional practices may be combined to form a scale that indexes teaching style. The standard for validating constructs and scales addresses the need to ensure that such measurement scales really do provide reliable and valid indices of the constructs they are designed to measure.

IEA studies are valuable only to the extent that the results are widely known and understood. One of the standards in this section relates to preparing a publication plan that will encourage widespread use of the study results, and includes guidelines for presenting findings clearly and accurately in a wide variety of IEA reports and publications. The issues of presentation discussed for particular reports cover organisation, format, facilitating correct interpretations of tabular and graphic material, and providing clear and accurate text. A closely related standard describes a review process to help ensure the quality of the publication plan and the primary reports issued from IEA studies. It specifies various critical steps in report development that require review. It also stresses that the process should be conducted according to a schedule prepared well in advance of the publication deadlines.

IEA aims for the widest possible use of the data from its studies, and promotes the release of data to researchers for secondary analysis as early as possible in the study schedule. The standard for releasing data discusses data release policies, and addresses issues that should be resolved so as to make data widely available. Allied to the need to make data available is the need to ensure that all study procedures are fully documented so that readers of international reports can evaluate the quality of the results, and secondary researchers have sufficient information to replicate analyses. The standard for preparing technical reports and documentation specifies what must be done so that an IEA study can achieve excellence in technical documentation.

Standard for Developing Sampling Weights

Purpose:

To assign sampling weights to all sampled units so that appropriate population inferences can be made.

Standard:

Sampling weights that accurately reflect the probability of selection should be assigned to each sampled unit.

Guidelines:

Consider the following guidelines when computing sampling weights:

a) *Sampling weights should have at least two components*: (i) a base weight that reflects the inverse of the selection probability of the sampled unit; and (ii) a correction factor for non-response that reflects the participation rate among sampled units in the stratum to which the sampled unit belongs.

b) *The non-response adjustment should reflect the difference between the achieved sample*, that is, the number of units (such as schools, classrooms or students) that actually participated, *and the expected sample size*, that is, the actual number of units that were originally sampled. For example, if only half of the sampled schools in a study participated, then a straightforward school non-response adjustment for the participating schools would be 2.0.

c) *In multi-stage samples, both the inverse of the selection probability and the non-response adjustment should be computed at each sampling stage*. A single overall weight that combines both weighting components from all sampling stages should be produced for each element in the sample.

d) *The sampling weight components should be derived from actual survey data and field documentation*, and should not be presumed to take *a priori* values based upon assumptions in the sampling design.

e) *The accuracy of the sampling weights* should be checked by producing population estimates of selected demographic variables (known as 'marker' variables) using the weighted sample data and comparing these to known population figures derived from census or other external sources. For example, the number of male and female students in a country at a grade level could be estimated from the weighted data and compared to known figures from official census data.

f) *The distribution of the weights among the sampled elements should be checked for the presence of unusually large weights.* Such outliers can contribute to sampling variance. Some consideration consequently should be given to moderating their influence (a process known as WEIGHT TRIMMING).

Standard for Reporting Sampling and Non-sampling Errors

Purpose:

To ensure that readers of IEA reports have all the information necessary to make informed decisions about the quality and reliability of the data used to generate the reports.

Standard:

Estimates of sampling error should be presented for all descriptive statistics included in IEA reports. Data quality measures on potential sources of non-sampling error, such as response rates, population coverage and exclusions, also should be reported.

Guidelines:

Take account of the following guidelines when reporting sampling and non-sampling errors:

a) *Sampling errors for descriptive statistics should be presented as either standard errors, coefficients of variation and/or confidence intervals.* Reports should include an explanation of how these statistics are to be interpreted.

b) *Sampling errors should be calculated so as to reflect the effects of complexities in the sampling design*. Standard statistical computer software (such as SAS and SPSS) uses algorithms intended for simple random samples, which generally underestimate the sampling errors in multi-stage samples. Software explicitly designed for estimating sampling error in complex survey data should be used. Software incorporating the jackknife algorithm can be readily developed for many purposes, and has been extensively used in the Third International Mathematics and Science Study, for example, with great success. Where appropriate, the effects of measurement error should be incorporated with the effects of sampling error.

c) *The data files and accompanying documentation should provide sufficient information* to permit the calculation of sampling errors for a wide range of estimates.

d) *Sampling participation rates should be calculated for each country at each stage of sampling, and also combined across all stages.*

e) Where possible, *quantitative information should be provided to indicate the potential for non-sampling error*. For example, measures of population coverage, exclusion rates and participation rates are useful indicators of possible response bias that should be reported, as are summaries of the extent of missing data, response inconsistencies among questionnaire items, and measurement error in scaled cognitive achievement scores.

Standard for Validating Constructs and Scales for Analysis

Purpose:

To ensure that all measurement scales (both cognitive and non-cognitive) used in analyses are valid measures of the intended constructs, so that consumers of study reports may have confidence in the validity of the results.

Standard:

Every effort should be made to establish the validity and reliability of all measurement scales reported in IEA studies. The content validity of measures of achievement should always be demonstrated. Similarly, some evidence of construct validity is also necessary for scales purporting to measure attitudes, opinions or other constructs. Where appropriate, evidence of the reliability of all scales should be provided in international reports.

Guidelines:

Take account of the following guidelines when reporting evidence of the validity and reliability of scales used in analyses:

a) *For measures of student achievement, the primary consideration will often be content validity*. This should be addressed by ensuring that the framework of the assessment has been certified by leading subject-matter experts as a satisfactory representation of the main content areas in the field, and that the test is an adequate measure of competence in these areas. IEA studies often incorporate measures of student 'Opportunity to Learn' (OTL), and such information may also provide useful information about the validity of achievement measures.

b) *Scales addressing attitudes and other constructs should also be reviewed by experts for content validity*. However, in addition, analyses should be undertaken to demonstrate their construct validity. Techniques such as factor analysis, cluster analysis or covariance structure analysis may be helpful in establishing relationships between variables hypothesised to comprise unified scales, for example.

c) Where scaling methods that make strong assumptions about the relationships among variables (such as item response theory) are used, *the plausibility of the assumptions should be investigated, and the fit of the data to the model should be assessed and reported*. The validity of the results may be jeopardised by poorly fitting models or violated assumptions.

d) *Small-scale field-test studies may be used to study the relationship between scale elements, and between proposed scales and other variables.* Evidence of convergent validity may be obtained when proposed scales correlate well with variables with which they should be related, and of divergent validity when they are unrelated to variables with which they should have no relationship.

e) *At a minimum, the reliability of measurement scales should be demonstrated in terms of an index of internal consistency*, such as Cronbach's Alpha. Other indices of reliability should also be calculated and reported, if possible. Scales with low reliability (for example, below 0.7) should be annotated in reports and interpreted with caution.

Standard for Presenting Study Findings

Purpose:

To facilitate widespread use of the study results and valid interpretation of the findings as they relate to the intended purposes of the study (see also Standard on Reviewing the Primary Reports of Study Findings).

Standard:

Here, the requirement is to ensure that the IEA study findings are presented clearly, accurately and in a timely fashion. To increase their use, study findings should be made available to different audiences in the forms most useful for those audiences. Note that even though this standard applies most directly to the presentation of study findings in publications sponsored directly by the IEA, plans for dissemination can also include other vehicles, such as presentations at professional meetings, papers for scientific journals and articles for professional magazines.

Guidelines:

To meet this standard, it will be necessary to:

a) *Specify the audiences for the publications and the needs of those audiences well in advance of the reporting deadlines.* There are many potential audiences for IEA results and reports, and those audiences will not want the same information delivered the same way. For example, some audiences may have an interest in the results, but not have the time to study them in detail. The findings from IEA studies therefore should be presented in different ways for different audiences. Modes of presentation can include, but not be confined to, technical reports, general audience reports, executive summaries, news releases, monographs, articles and brochures.

b) *Develop a publication and reporting plan that addresses the needs of as many audiences as possible within the available resources*. The size and detail of the plan will vary with the size and complexity of the study, but it should include the types and numbers of publications to be produced, the process for developing and disseminating the publications, and the production schedule. Schedule the analysis and reporting activities to ensure timely availability of the publications containing the study results and make efforts to maintain that schedule.

c) *Ensure that the audience is not confused by other (albeit interesting) information that is not vital for their needs.*

d) *Determine the achievement test and questionnaire results that will be presented*, and ensure that they are accurately recorded in tables, graphs or plots, with tests of statistical significance where appropriate. This task includes:

- Choosing a measurement scale for achievement results that is consistent with the purposes of the test. When raw scores are used, they should be demonstrated to have the same desirable properties as do scaled scores.
- Presenting standard errors or confidence intervals for the statistics, either in the same table as the data being presented or in a separate table.
- Providing row and column totals where possible.
- Checking for consistency between and within tables.

e) *Ensure that tables and graphs display data in such a way that readers can conveniently evaluate the accuracy of the results presented.* This may include grouping countries according to different characteristics at different times, such as particular attributes of their educational systems, or the degree of their sample coverage.

f) *Make sure that the data in each table and graph are correctly identified and labelled* (each table and graph should be able to stand alone):

- Identify the table as from an IEA study (e.g., include the IEA logo).
- Identify the name and date of the IEA study.
- Include a short title that concisely states the subject of the table.
- Provide labels for the names of variables and for their categories.
- Provide source notes. When the data presented come from multiple sources or from a source that is not the direct subject of the report, the source note must clearly match the data items with the source.
- Use footnotes that explain special cases for the numbers used in calculations, and for when rounded results are presented that initially appear inconsistent.

g) *For graphic displays, make sure that the presentation enhances the ease and accuracy of data interpretation*. Keep in mind:

- Simplicity and clarity. Cluttered graphs are ineffective.
- Clear labels for the horizontal and vertical axes. When using time series intervals, the horizontal axis intervals should be equal to the time intervals. Ordinarily, the vertical scale should start

at zero. Otherwise, scale breaks should be clearly visible. Labels should be placed on the tick marks for axes.

- Avoid stacked bar graphs, stacked area line graphs, and histograms. These are difficult to read and easily misinterpreted.

h) *Ensure that the tables and graphs can be reproduced* using widely available photocopying techniques without losing quality in terms of accuracy, interpretability and legibility.

i) *Organise the publication in a way that increases ease of use*. Highlight the most important points by putting the most important information first, using descriptive section headings and designing the spacing and layout so that it improves readability. For some publications, this also may involve providing an Executive Summary or other easily accessible summaries of the study implementation and findings.

j) *Make sure that the text of the publication is readable*. Use active verbs as much as possible, keep the paragraphs short, vary the format, and try to personalise the text as much as possible. One way to do this is to keep potential readers of the report in mind.

k) *Use tables, graphs and figures as a way of presenting information efficiently and effectively*. Tables, graphs and figures can be used to present information as well as data, including steps in program implementation, aspects of sample attrition, and domain coverage of the tests and questionnaires.

l) *Provide summary and reference information about the implementation of the study* so that the user can evaluate the appropriateness and technical adequacy of the study results. This information includes the:

- purpose of the study
- mode of administration
- sample coverage
- coverage of the framework
- types of items used
- translation verification procedures
- data collection procedures
- scoring procedures
- scaling methods
- analysis procedures
- staff involved in the study

(see also Standard for Technical Reports and Documentation).

m) *Design each publication in an attractive format that suits the needs of the readers.* If possible, involve a graphics specialist in the reporting plans. This can help improve the presentation of individual tables and graphs, the highlighting of important information, and the overall layout of the publication.

n) *Have each publication edited by a professional.* This person can look for unnecessary and confusing words and phrases as well as correct usage, spelling and punctuation.

Standard for Reviewing the Primary Reports of Study Findings

Purpose:

To ensure that the primary reports of study findings present the most important findings accurately and in a way that is consistent with the aims of the study.

Standard:

The publication plan for the study as well as primary reports of study findings must be thoroughly reviewed by the appropriate constituencies and advisory committees before publication. These reviews should occur at critical points during report development and should comprise, for example, the report outline, shells for the proposed tables and graphs, drafts of the text and the data for tables and graphs, and the final draft of the complete report.

Depending on the resources available for the study, the reviews should be conducted in meetings of the relevant parties so that differing points of view can be resolved in the best way possible. However, supplementary tele-conferencing and mail reviews also will be necessary. All steps of the review process should be conducted in accordance with a schedule prepared well in advance of the publication deadline, and all parties involved in the review should be given adequate notification about the review schedule.

Guidelines:

To meet this standard, the international co-ordinating centre (ICC) or responsible party should use the following process:

a) *Have the publication plan for the study reviewed by the various advisory committees associated with the study*, and revise the plan accordingly. Depending on the study, these committees generally include the Steering Committee, the Subject Area Advisory Committee, the Technical Advisory Committee and the Publications and Editorial Committee.

b) *Have the publication plan for the study reviewed by representatives of the national research centres participating in the study*, and revise the plan accordingly.

c) *Have the reporting plans reviewed by the IEA Technical Advisory Committee and the IEA Publications and Editorial Committee*, and revise the plans accordingly.

d) For the primary reports described in the publication plan, *ensure that the report outlines, analyses procedures and proposed shells for the tables and graphs for the reports are reviewed by the various advisory*

committees associated with the study and by the representatives from the national centres participating in the study. The emphasis is on ensuring that each report includes the most important information for its purpose, is as brief and concise as possible in light of this purpose, and that the presentations are technically accurate. Revise the plans for the reports accordingly. These reviews should be held well in advance of the publication deadlines for the reports (at least one year is recommended).

e) *Have representatives of the national research centres review the data for their countries*. This work is often done by distributing almanacs of the study results electronically and in hard copy. These almanacs should be all inclusive, including results for the cognitive items and background questionnaire results. The goal is to ensure accurate data for each country before analysis and incorporation of the data into the table shells and graphs.

f) *Ensure the ICC reviews the cognitive and questionnaire data across all countries for consistency and accuracy* before analysis and incorporation of the data into the table shells and graphs.

g) *Before writing text associated with the tables and graphs, conduct the appropriate analyses and incorporate the data into the tables and graphs*. Using an electronic procedure for this process will help ensure accuracy. Staff at the ICC should review the data in the tables and graphs for accuracy in terms of the analyses required, credibility of results, and consistency within and across these items.

h) *Have representatives of the national research centres review the draft reports*, including the data in the tables and graphs and the associated text. The emphasis should be on clarity and accuracy. The goal of completeness must be balanced against the goals of accuracy and timeliness. Revise the report organisation, tables and graphs and the text in accordance with the review.

i) *Have the final draft of the reports undergo a thorough review by a professional editor*. This work should include attention to headings and labels for tables and graphs as well as to the textual matters of clarity, brevity, usage, grammar, spelling and punctuation. Make the appropriate revisions.

j) *Have the final drafts of the reports reviewed by representatives of the national research centres, the editorial committee for the study, and the IEA Publications and Editorial Committee*. Revise the reports in accordance with this review.

k) *Ensure the ICC proofs all aspects of the final reports before preparing them for the printing process.*

l) *Ensure the ICC proofs all aspects of pre-print copies of the reports before printing multiple copies of the reports.*

m) *Ensure the ICC checks the quality of the printed copies of the reports before distribution*. This work should include a check on the quality of the printing, that all pages are included in the reports, and that the pages are in the right order and facing the right direction.

Standard for Releasing Data

Purpose:

To ensure that an international database is available under pre-defined conditions.

Standard:

A data release policy should be decided at the beginning of the study. Agreement with this policy should be secured in writing from all participants at an early stage. Participants should also agree on a timetable that defines when and for whom the data are available and whether a distinction should be made between users who are from institutions involved with IEA and users from other institutions. At the end of the study, the released data should be made available to all interested users.

Guidelines:

To meet this standard:

a) *Prepare a release policy defining who will get what data when and in which format.* All participating country representatives need to reach agreement on this policy at an early stage of the study.

b) *Ensure that participants are given the opportunity to discuss and agree on timelines* for the data release and the format for the data release.

c) *Distribute the data in a format that is accessible to the widest possible audience*. Avoid reliance on data formats that require proprietary software for access.

d) *Fully document all data at the time of release*. Such documentation must be comprehensive enough to permit a user without previous familiarity with the study to analyse the data (see also Standard for Documenting Data-processing Activities).

Checklist: ☑

- ☐ Ensure that all data products are fully documented.
- ☐ Secure written permission for data release from all participants in time for the release.

Standard for Preparing Technical Reports and Documentation

Purpose:

To ensure that the study design, instrument development, data collection, and analysis and reporting procedures are described in sufficient detail to permit them to be evaluated and replicated.

Standard:

All aspects of the study design, development, operations and analysis should be clearly documented in a timely manner. Reports intended for a general audience should include a brief technical appendix with an explanation of the major technical features of the study. Detailed technical documentation aimed at a technically sophisticated audience should also be produced, ideally in a separate technical report. This document should include specifications of all design, procedural and analytic aspects of the study.

Guidelines:

To meet the requirements of this standard, ensure that:

a) *The technical reports describe in detail the study design*, including instrument development and data collection, processing and analysis procedures.

b) *The technical reports describe the actual procedures used to collect, process and analyse data*. The reports should also describe the procedures used to ensure that the data were collected in a consistent manner in each participating country.

c) *The technical documentation is produced in a timely manner*. In a complex study extending over many years, it may be desirable to produce technical volumes at critical points of the study. For example, the Third International Mathematics and Science Study produced an initial volume (describing the design and development of the study) while data were being collected, and it published subsequent volumes (documenting the analysis and reporting) at a later stage (see Martin & Kelly, 1996, 1997, 1998).

d) *Discrepancies between the design and the actual procedures used in collecting, processing or analysing data are described and justified* in technical reports.

e) *The implementation of the sampling design in each country is described*, and accurate information is included on population coverage, participation rates and the sampling precision achieved. Countries deviating from study standards for coverage or participation should be clearly annotated.

f) *The reliability and validity of data collection instruments are documented* in technical reports.

g) *The technical reports clearly document the quality assurance procedures used in all stages of collecting and processing the data.* Particular attention should be paid to procedures for monitoring the administration of achievement tests or questionnaires for students.

h) *Procedures for coding or quantifying responses to tests or questionnaires are described,* and the steps taken to ensure accuracy and consistency across raters are documented.

i) *Technical documentation is provided on the processing and storage of data,* including test and questionnaire protocols, scoring and coding manuals, data dictionaries, and specific information on the re-coding of variables or the construction of derived variables for analysis and reporting.

j) *Advanced or innovative techniques for scaling or analysing study data are documented,* and evidence of their efficacy is provided if such information is not already available in the public domain.

k) *At the end of the study, the data is made available for secondary analysis.* The data should be fully documented and in a readily transportable, machine-readable format.

l) *The documentation of analytic procedures is sufficiently detailed to permit someone with access to the data to replicate the results.*

Bibliography

Burstein, L (Ed) (1992) *The IEA Study of Mathematics III: student growth and classroom processes*. Oxford: Pergamon Press.

Committee to Develop Standards for Educational and Psychological Testing of the American Educational Research Association, The American Psychological Association, and The National Council on Measurement in Education (1985) *Standards for Educational and Psychological Testing*. Princeton, NJ: The American Psychological Association.

Converse, J M and Presser, S (1986) *Survey Questions: handcrafting the standardised questionnaire.* Newbury Park, CA: Sage Publications Inc.

Educational Testing Service (1987a) *ETS Standards for Quality and Fairness*. Princeton, NJ: Author.

Educational Testing Service (1987b*) Guidelines for Developing and Scoring Free-Response Tests*. Princeton, NJ: Author.

Elley, W (1992) *How in the World Do Students Read*? Amsterdam: International Association for the Evaluation of Educational Achievement.

Elley, W (1994) *The IEA Study of Reading Literacy: achievement and instruction in thirty-two school systems*. Oxford: Pergamon Press.

Joint Committee on Testing Practices (1988) *Code of Fair Testing Practices in Education.* Washington, DC: National Council on Measurement in Education.

Lyons Morris, L, Taylor Fitz-Gibbon, C and Freeman, M E (1987) *How to Communicate Evaluation Findings*. Thousand Oaks, CA: Sage Publications Inc.

Martin, M O and Kelly, D L (Eds) (1996) *Third International Mathematics and Science Study Technical Report, Volume I: design and development.* Chestnut Hill, MA: Boston College.

Martin, M O and Kelly, D L (Eds) (1997) *Third International Mathematics and Science Study Technical Report, Volume II: analysis and reporting—Populations 1 and 2.* Chestnut Hill, MA: Boston College.

Martin, M O and Kelly, D L (Eds) (1998) *Third International Mathematics and Science Study Technical Report, Volume III: analysis and reporting—Population 3.* Chestnut Hill, MA: Boston College.

Martin, M O and Mullis, I V S (Eds) (1996) *Third International Mathematics and Science Study: quality assurance in data collection.* Chestnut Hill, MA: Boston College.

National Center for Education Statistics (1992) *NCES Statistical Standards.* Washington, DC: U.S. Department of Education Office of Educational Research and Improvement.

Robitaille, D F and Garden, R A (1996) *Research Questions and Study Design, TIMSS Monograph No. 2.* Vancouver, BC: Pacific Educational Press.

Robitaille, D F, Schmidt, W H, Raizen, S, McKnight, C, Britton, C and Nicol, C (1993) *Curriculum Frameworks for Mathematics and Science, TIMSS Monograph No. 1*. Vancouver, BC: Pacific Educational Press.

Salant, P and Dillman, D A (1994) *How to Conduct Your Own Survey*. New York, NY: John Wiley and Sons Inc.

Westat (1991) *SEDCAR (Standards for Education Data Collection and Reporting)*. Washington, DC: US Department of Education.

International Association for
the Evaluation of Educational
Achievement