



IEA

Researching education, improving learning

IEA's Technical Standards for International Large-Scale Assessment

Editors:

Michael O Martin

Julian Fraillon

Heiko Sibberns

Michael O Martin
TIMSS & PIRLS International
Study Center Boston College,
United States (retired)

Julian Fraillon
IEA Amsterdam
The Netherlands

Heiko Sibbers
IEA Hamburg
Germany

With contributions from

Betina Borisova
Ekaterina Buzkich
David Ebbs
Eugenio Gonzalez
Seamus Hegarty

Sabine Meinck
Sebastian Meyer
Irini Moustaki
Lauren Musu
Keith Rust

Ulrich Sievers
Matthias von Davier
Kentaro Yamamoto

IEA Secretariat
Keizersgracht 311
1016 EE Amsterdam
The Netherlands
Tel.: +31 (0)20 625 3625
Fax: +31 20 420 7136
secretariat@iea.nl



The International Association for the Evaluation of Educational Achievement (IEA), with headquarters in Amsterdam, is an independent, international cooperative of national research institutions and governmental research agencies. It conducts large-scale comparative studies of educational achievement and other aspects of education, with the aim of gaining in-depth understanding of the effects of policies and practices within and across systems of education.

© Stichting IEA Secretariaat Nederland 2025. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

This work is subject to copyright. All commercial rights are reserved by the author(s), whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Regarding these commercial rights a non-exclusive license has been granted to the publisher.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

ISBN 9789079549368

Michael O Martin • Julian Fraillon •
Heiko Sibberns (Eds.)

IEA's Technical Standards for International-Large Scale Assessments

Betina Borisova • Ekaterina Buzkic • David Ebbs •
Eugenio Gonzalez • Seamus Hegarty • Sabine Meinck •
Sebastian Meyer • Irini Moustaki • Lauren Musu •
Keith Rust • Ulrich Sievers • Matthias von Davier •
Kentaro Yamamoto

Contents

Foreword	6
Introduction	7
Background.....	7
Characteristics of IEA's International Large Scale Assessment Studies	8
Structure of the Standards.....	10
Standards for Designing, Managing, and Implementing IEA Assessments	11
Section Overview.....	11
Standard for Initial Planning of an IEA Study	12
Standard for Establishing an International Study Center.....	13
Standard for Formulating and Refining Study Aims and Questions.....	14
Standard for Designing an IEA Study.....	15
Standard for Choosing a Computer-Based Assessment System	16
Standard for Ensuring Accessibility Support	17
Standard for Data Protection	18
Standard for Developing a Sampling Plan.....	19
Standard for Choosing Data Collection Methods	20
Standard for Developing a Quality Assurance Plan.....	21
Standard for Preparing an Analysis Plan	22
Standard for Selecting Psychometric and Statistical Analysis Procedures.....	23
Standards for Developing Data Collection Instruments.....	24
Section Overview.....	24
Standard for Developing Assessment Frameworks	26
Standard for Developing Specifications for Achievement and Questionnaire Instruments.....	27
Standard for Achievement and Questionnaire Item Development	28
Standard for Developing Scoring Guides for Human Scoring	30
Standard for Translations and for Verifying Translations.....	31
Standard for Field Testing Data Collection Instruments and Procedures.....	32
Standards for Data Collection and Processing.....	34
Section Overview.....	34
Standard for Drawing a Sample	35
Standard for Planning for Data Collection.....	37
Standard for Minimizing Response Burden.....	39

Standard for Selecting and Training Data Collection Staff	40
Standard for Implementing Data Collection Quality Control Procedures.....	41
Standard for Documenting National Data Collection	42
Standard for Planning Data Requirements, Data Saving, and Data Processing	43
Standard for Planning for Data Preparation and Processing in Paper-based Assessments	44
Standard for Human Scoring of Student Responses	45
Standard for Machine Scoring of Student Responses	46
Standard for Processing and Checking Data	47
Standard for Documenting Data-Processing Activities.....	48
Standards for Analyzing Data and Reporting Results	49
Section Overview.....	49
Standard for Conducting Psychometric and Statistical Analyses.....	50
Standard for Developing Sampling Weights.....	51
Standard for Reporting Sampling and Non-sampling Errors.....	52
Standard for Presenting Study Findings.....	54
Standard for Reviewing the Primary Reports of Study Findings.....	55
Standard for Releasing Data	56
Standard for Preparing Technical Reports and Documentation	57
Standard for Ensuring Data Integrity	58
Standard on Training Primary and Secondary Users of IEA Data.....	59
References	60
Further Reading.....	61

Foreword

IEA (International Association for the Evaluation of Educational Achievement) strives to improve teaching and learning across educational systems through high-quality international large-scale assessments, innovative research, expert guidance, and global collaboration. The work conducted by our organization is bolstered by the international cooperation of national research institutions, governmental research agencies, scholars, and analysts who are committed to advancing educational knowledge and practice. By conducting large-scale comparative studies of education worldwide, IEA provides educators, policymakers, and parents with crucial insights into student performance, with a mission to understand educational practices, processes, and policies.

All IEA studies are developed thoroughly and in collaboration with our network, ensuring that each study remains grounded in curriculum-based frameworks reflective of diverse educational systems. The initial standards, published in 1999, aimed to consolidate the best practices and methodological rigor in IEA's approach to educational assessment, addressing the unique needs of international studies. Since 1999, many changes to IEA studies can be seen. All IEA studies are now computer-based assessments, which offer opportunities but also new challenges. Furthermore, from 1999 onwards, all IEA studies became cyclical studies that are conducted regularly to measure trends in education. This publication of IEA's technical standards builds upon the foundation laid by its predecessor while also addressing key updates to guide future IEA assessments. This allows for IEA to adapt to new methodologies while upholding the tradition of meticulous standards that its studies are known for.

Our editors, Michael O. Martin, Julian Fraillon, and Heiko Sibberns, have ensured that these updates honor IEA's tradition of collaboration and high standards while pushing our commitments to innovation forward. We further express our gratitude to the contributors—Betina Borisova, Ekaterina Buzkich, David Ebbs, Eugenio Gonzalez, Seamus Hegarty, Sabine Meinck, Sebastian Meyer, Irini Moustaki, Lauren Musu, Keith Rust, Ulrich Sievers, Matthias von Davier, and Kentaro Yamamoto—whose well-rounded expertise and dedication have been essential to furthering this publication's quality.

IEA assessments, such as TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), ICCS (International Civic and Citizenship Education Study), and ICILS (International Computer and Information Literacy Study), are recognized as the gold standard in international large-scale educational studies; they provide reliable, high-quality data that inform educational policy and practice worldwide. These standards serve not only to ensure technical consistency but also to enable the comparability and integrity of data that stakeholders rely upon.

In an era of increasingly complex educational challenges, these standards reinforce IEA's commitment to quality and innovation. Thank you to everyone who contributed their time and knowledge to updating the technical standards of IEA. This publication is a valuable resource to support future studies in continuing to yield insights that are both scientifically sound and relevant to educational research worldwide.

Dirk Hastedt
Executive Director IEA

Introduction

Background

In the 25 years since the International Association for the Evaluation of Educational Achievement (IEA) first published its technical standards (Martin et al., 1999), much has changed in the world of international large-scale assessment (ILSA), as the field has come to be known. There has been significant growth in such assessments, and corresponding increase in the number of countries participating in them (Hernández-Torrano & Courtney, 2021; Addey, et al., 2017). Accompanying this growth has been greater visibility of assessment results among the public and greater reliance on assessment data by policy makers and educators. In response to the rapid expansion of the field and increased importance of assessment data, IEA's Technical Standards for International Large Scale Assessment is a timely review, extension, and update of the original Technical Standards for IEA Studies published in 1999, reflecting also the increasing use of the ILSA terminology in place of the more generic "IES Studies."

As the organization that pioneered international comparative studies of student achievement in the 1960s, IEA has always aspired to the highest technical and professional standards. These Technical Standards seek to document the critical activities involved in carrying out IEA's international large-scale assessments and make explicit the criteria against which such activities should be judged. The Technical Standards are intended to serve as guides to good practice for future IEA assessments, and as a yardstick against which such assessments may be evaluated by policy makers, researchers, and other interested parties. The work of drafting and reviewing the Technical Standards was conducted under the auspices of the IEA Technical Executive Group.

Although the principles underlying the standards and guidelines laid out in the first edition of the Standards are still relevant today, there have been several major developments that had to be considered when updating the standards. These include:

- **Evolution of the major IEA studies from one-off studies to regularly conducted programs of assessment measuring trends**

Since the 1999 Standards were developed, TIMSS¹ and PIRLS² have become ongoing programs, with TIMSS assessing trends in student achievement in mathematics and science at fourth and eighth grades every four years, and PIRLS assessing trends in reading literacy at fourth grade on a five-year cycle. ICCS³ and ICILS⁴ also are now established programs. ICCS assesses civics and citizenship education at eighth grade on a regular basis. ICILS assesses computer and information literacy at eighth grade on a five-year cycle.

- **Transition from print-based to computer-based assessment**

The enormous growth in digitalization in recent years has impacted the design and implementation of IEA's studies in many ways, not the least of which has been the transition of its assessments from print-based delivery to computer-based assessment (CBA). These changes need to be reflected in updates to the standards, which were originally developed for a print-based delivery environment. The transition to CBA impacts study design, instrument design, instrument preparation, data collection, and analysis. In addition to CBA necessitating changes in some pre-existing methods, it affords new opportunities such as the collection of process data from respondents, the integration of AI into study design and implementation, and the implementation of adaptive delivery systems.

- **Increased study complexity**

IEA's studies have increased in complexity in recent years, requiring additional expertise and coordination efforts from study centers and project staff. These include moving to group adaptive designs in TIMSS and PIRLS, adding longitudinal options to TIMSS, offering the opportunity to administer two studies

¹ Trends in International Mathematics and Science Study

² Progress in International Reading Literacy Study

³ International Civic and Citizenship Education Study

⁴ International Computer and Information Literacy Study

together (TIMSS and PIRLS in 2011) or in the same school year (TIMSS and ICILS in 2023), including optional modules such as computational thinking in ICILS, and offering regional modules in ICCS.

- **Greater emphasis on data protection**

Increased concerns about rights to privacy and data protection have led to greatly increased data protection regulations, most notably the European Union's General Data Protection Regulation (GDPR), which has had far-reaching effects on data collection, processing, and dissemination activities.

IEA has both grown as an organization and become more directly involved in study activities since the publication of the 1999 standards. IEA Amsterdam, previously known as the IEA Secretariat, is responsible for funding and membership issues, and coordinates translation verification of data collection instruments for all studies. It also hires and trains quality control monitors to monitor and report on data collection activities. IEA Hamburg, previously known as the IEA Data Processing and Research Center (DPC), continues to be responsible for data processing and cleaning, database construction and, depending on the study, sampling, and study coordination. In some studies, it also contributes to research analyses and reporting.

In addition to its international large-scale assessments (TIMSS, PIRLS, ICCS, and ICILS), IEA has, over the past 20 years, conducted national studies in Germany in cooperation with a several scientific consortia, research institutes, and universities. IEA also implements a wide range of studies on behalf of international partners, including the OECD (TALIS, TALIS Starting Strong, and PIAAC), the Inter-American Development Bank (PRIDI), and the UNESCO Institute for Statistics (Rosetta Stone and REDS). Although IEA seeks to achieve the highest standards in all of its endeavors, the Technical Standards are intended specifically for its international large-scale assessments.

Characteristics of IEA's International Large Scale Assessment Studies

The Technical Standards acknowledge explicitly the co-operative nature of IEA, and its project structure and management style, and outline the major characteristics of its international large-scale assessments, also known for brevity as IEA studies. The standards are applicable at both the international level and at the national level.

At the international level, each of the major international large-scale assessments has an international study center⁵ that is responsible for all aspects of project design and implementation. This includes establishing the study aims and plan, defining the survey population(s), establishing the sampling and instrument designs, developing and updating assessment frameworks, developing and field-testing assessment and questionnaire instruments, implementing and documenting data collection procedures, scoring and quality assurance activities. Data processing activities at the international level focus on checking the quality of the data, constructing measurement scales, and conducting analyses for international reports. Drafting and reviewing international reports and disseminating results are also responsibilities of the international study center in collaboration with IEA.

At the national level, IEA national centers are responsible for funding and staffing project activities, and for appointing a national research coordinator responsible for conducting the study nationally in accordance with internationally agreed study parameters. In particular, the national research coordinator is responsible for:

- reviewing draft assessment framework documents
- working with international sampling consultants to confirm sampling plans and sampling frames that meet international specifications
- participating in assessment instrument development, including suggesting potential topics and content for inclusion in instruments, and providing feedback on draft instruments during the instrument development

- preparing instruments for use in data collection (i.e., adapting, translating, and responding to verification feedback) of instruments for use in data collection
- preparation of procedural manuals (provided by the international study center) for use in national data collection
- selecting and providing the appropriate technical infrastructure to support efficient and reliable data collection from all respondents
- conducting the field test and main data collections
- scoring student responses
- capturing and providing national data in the format defined by the international study center
- conducting national analyses
- producing national reports as required by national authorities.

National centers may choose to outsource certain project activities, such as instrument translation, assessment administration, or student response scoring, to external contractors. In these situations, the national research coordinator is responsible for ensuring that the contractor adheres to project specifications and guidelines.

IEA's international large-scale assessments measure variations in student achievement over time (trends) in one or more curricular areas, including mathematics, science, reading, information technology, and civic education. Student achievement is measured by administering assessment items, developed as part of the study, to random samples of students selected to represent national populations. IEA's approach to measuring trends in student achievement in a changing world is described in Mullis & Martin (2022). Information about the students' backgrounds, attitudes, and interests is collected by means of self-report questionnaires administered to the students at the same time as the achievement items. This information can be used to describe aspects of students' characteristics and contexts that may relate to variations in student achievement. In some cases, this information may be used to measure curriculum outcomes that are concomitant to achievement. Information about the schools the students attend is collected by one or more questionnaires addressed to senior staff in the school (usually including the school principal). Information about classroom practices and teachers' experiences and attitudes is collected through a questionnaire to teachers of the target grade (that may or may not include the classroom teacher). Information about the curriculum and the education system (e.g., the structure of the system, national and regional policies, locus of decision-making) is provided through questionnaires completed by national research coordinators.

The following characteristics of IEA assessments have helped shape these Technical Standards.

- *Exemplary quality:* IEA assessments aim to collect valid and reliable data that are comparable across participating countries using methods and procedures that provide contemporary models of best practice in large-scale assessment.
- *Continuous improvement:* IEA assessments adopt processes of continuous improvement in which internal reviews of all aspects of the studies are reviewed and, where appropriate, improvements are implemented. New and emerging methods for field operations, quality assurance, psychometrics, statistical analyses, and reporting are considered and, where appropriate, adopted for use.
- *Ongoing development:* The development and updating of the research themes, assessment frameworks, and assessment and questionnaire instruments is a major component of IEA studies. This is done with the aim of maintaining contemporary relevance for each study, while still enabling trends to be reported.
- *Complex assessment designs:* The breadth and range of difficulty of achievement items necessary to cover the curriculum learning area content (as defined in the assessment framework) is typically too large to accommodate in a single test form. Consequently, student test instruments typically comprise large pools of items that are distributed across many test forms. This necessitates both complex assessment

⁵ The international study center for TIMSS and PIRLS is at Boston College. The Australian Council for Educational Research was the international study center for the first three cycles of ICCS and the first two cycles of ICILS. The University of Bath is the international study center for the fourth cycle of ICCS, and IEA is the international study center for the third and fourth cycles of ICILS.

designs to balance the test content across forms, and appropriate complex psychometric procedures to analyze and report assessment results.

- *Measuring trends:* Measuring trends in student learning outcomes⁶ associated with school curriculum areas is a fundamental objective of most IEA studies. Trend measurement necessitates that a substantial number of common items with appropriate psychometric characteristics be held secure and included in successive cycles of each study.
- *Detailed measures of learning contexts:* Contextual information is collected by means of questionnaires from the students, their teachers, and school personnel (usually from the school principal or delegate and potentially also curriculum leaders). There also may be a questionnaire collecting national-level curriculum- and policy-related information to be completed by the national center. The development of these questionnaires is usually also a major component of the study.
- *Multistage sampling:* Data are collected by means of sample surveys conducted in the school setting. Data-collection methods are designed to make efficient use of the school and classroom organization. Sampling plans make use of multistage designs with schools, classrooms, and, as appropriate, students or teachers as stages in the sampling process, and rely on school personnel such as principals, administrators, and teachers to provide data. Over the years, data collection has evolved from print-based to computer-based/online methods.
- *Cross-national data comparability:* Data are collected simultaneously in large numbers of countries by national teams using internationally agreed instruments and following internationally agreed procedures. These are designed to ensure data comparability across countries. Item adjudication and scaling analyses include empirical measures of the cross-national consistency of the items and scales used to measure the constructs reported in the studies.
- *Flexibility to consider and, where appropriate, accommodate national needs:* Translation and adaptation of the international instruments and procedures to local conditions is conducted by the national team in each country. National teams may add extra achievement or questionnaire items as national options, provided these options do not compromise the quality of international instruments. The inclusion of national options takes place only with the express approval of the international study center. National teams are responsible for quality control of all national activities.
- *Central coordination by an international study center:* Each study is directed and coordinated by an international study center, which establishes the aims of the study, leads the development of data collection instruments and procedures through a co-operative process, and ensures high quality in all aspects of the project, including sampling, data collection, and data processing, statistical and psychometric analysis, and reporting of results.

Structure of the Standards

IEA's technical standards focus primarily on the design and management of assessments at the international level, and on those aspects of the national implementation of studies that are necessary to ensure the timely collection of high-quality internationally comparable data. Accordingly, many of the standards in this document are of primary relevance to staff of the international study center, whereas others apply to staff in national centers. Each standard identifies the intended audience, as necessary. The standards are grouped into four areas:

1. Design, management, operation, and quality control of international studies.
2. Construction of instruments for measuring student achievement and questionnaires for collecting background information from students, teachers, and schools.
3. Data collection through sample surveys in the school setting.
4. Data processing, statistical and psychometric analysis, and reporting.

Each standard statement includes the purpose of the standard, the standard itself, and a set of suggested guidelines for meeting the standards.

⁶ Learning outcomes typically comprise student achievement. However, for selected curricular areas they may also include attitudes, behaviors, and behavioral intentions.

Standards for Designing, Managing, and Implementing IEA Assessments

Section Overview

IEA's international large-scale assessments are complex, high-profile profile, and costly endeavors. It is therefore essential to pay attention at the outset to the study's aims and objectives. Time and energy spent in the planning stages on establishing consensus on aims, objectives, methods, and standards will be repaid many times over in terms of efficiency of operations during the lifetime of the study. This applies to new studies and to new cycles of existing assessments. The study's initial plan for an IEA large-scale assessment should be a concise summary of the objectives, methods, schedule, costs, and benefits of the projected study. It should contain sufficient information for prospective participants and intended funding agencies to evaluate the scope of the study, as well as the projected costs and benefits.

Given that international studies are crucially dependent on the quality of international direction and coordination, an effective international study center is essential to the success of any IEA study. The international study center 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 underway, 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. The design should specify how the study questions will be addressed and provide a detailed plan for conducting the study. This should include specifications for data collection methods, including computer-based assessment systems as required, as well as provision for accessibility support to maximize inclusion and measures to ensure compliance with all data protection requirements.

Satisfactory implementation of the design will require the development of a sampling plan that specifies the sampling procedures that are appropriate for the study and the standards for sampling precision that are required. Since the administration of achievement items and questionnaires is a central feature of IEA assessments, a data collection procedure that capitalizes on efficiencies of information and communications technologies under standardized 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 assessment.

An analysis plan that spells out how the data will be analyzed to achieve the study aims is an essential component of the planning process. This should outline the state-of-the-art psychometric and analytic procedures to be applied. 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 developing an initial proposal for an IEA study.

Standard:

The initial proposal for an IEA study should contain sufficient information to enable intending participants to evaluate its scope and projected costs and benefits. This information should include the rationale for the study, its goals and objectives, and the methodology that will be used, including the technological requirements, target populations, analytic methods, implementation schedule, and preliminary cost and staffing estimates.

Guidelines:

The initial planning for an IEA study should include:

- A justification of the study in terms of the issues and questions to be addressed, including the relation of those issues and questions to research literature, stated in terms of its specific goals and objectives.
- A preliminary study design that describes the assessment design, target populations, sample design, instrument development, data collection methods and technology, and the methodological and technical issues to be resolved. The design should include requirements for acceptable population coverage and response rates.
- 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 participating country. The study plan should include consideration of accessibility support for participants and make provision for achievement item security and compliance with data protection regulations.
- A preliminary analysis plan that identifies analysis issues, objectives, major variables, and reporting scales, as well as proposed psychometric and statistical techniques.
- A preliminary publication and dissemination plan that identifies proposed major publications and their target audiences.
- A proposed 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 coordinators, and which are a joint responsibility.
- Preliminary international cost estimates, broken down by major project tasks, for staff (in terms of person/months or weeks) and for all other costs. There should be sufficient information to enable intending participants to estimate their national costs.

Standard for Establishing an International Study Center

Purpose:

To ensure that the international study center is properly staffed, equipped, and funded to direct all aspects of study design, development, implementation, analysis, and reporting.

Standard:

As the international management and coordination center for the study, the international study center should be well connected to academic research related to the domain under study, have staff with a solid research and methodology background, and the technical and managerial competence and experience necessary to conduct the study successfully. It also should have the infrastructure to support the study effectively, including accommodation, support staff, data processing equipment, and information and communications facilities.

Guidelines:

To meet the challenge of directing and coordinating an IEA study, the international study center should have:

- Senior international study center staff with expertise in large scale educational research covering learning content domains, research methodology, analytic methods, and a good understanding of current academic and policy research in education. More specifically, staff should have:
 - the vision, leadership skills, technical knowledge, and expertise to conceptualize and direct all aspects of the design and operation of the study
 - a thorough understanding of the curricular areas to be studied (for example, mathematics, civic engagement)
 - the managerial competence, and experience and communication skills to set and maintain schedules, hire and motivate staff, develop and manage budgets, disseminate information about the study, and maintain communications between study partners
 - Fluency in English, so that all documents produced by the international study center meet the highest semantic and syntactic standards.
- Suitable office accommodation, adequate working and meeting space, and access to the necessary technological and computational resources for the various phases of the study, including modern information and communications technology.

When making the transition between international study centers, all study documentation, as the property of IEA, should be made readily available to the new international study center as soon as possible, ensuring staff are fully informed about how the study was conducted and have all necessary materials to continue the work and maintain communication with national research coordinators.

Standard for Formulating and Refining Study Aims and Questions

Purpose:

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

Standard:

The study's aims and questions should address issues of central importance to participating countries, relate to published research literature, be clear in their meaning, implications, and assumptions, and be answerable from the data to be collected by the study.

Guidelines:

To meet this standard:

- Policymakers, educational practitioners, funding agencies from participating countries, and international sponsoring organizations should have an early opportunity to review the proposed study aims and questions and suggest appropriate amendments.
- Research Questions should reflect knowledge of the relevant academic research literature and be formulated so that they can be answered empirically through practical data collection activities.
- The study's methodological approach (e.g., one-time cross-sectional survey, repeated cross-sectional survey, longitudinal study, etc.) should be specified so that reviewers can evaluate its appropriateness to answering the research questions.
- Each study question should be operationalized in terms of variables that can be validly and reliably measured in the languages of the participating countries.

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 clearly show how the study questions are to be addressed and provide a plan for the collection, processing, and analysis of data, and the reporting of results.

Guidelines:

- The study design should prescribe the study's methodological approach and show how it will address the study research questions. It should also explain the kinds of data collected and the inferences that can be made from them based on the planned design. The design should include a schedule for the main study activities, making allowance for countries on northern and southern hemisphere timelines.
- The study design should include an assessment or conceptual framework specifying the information to be collected and the constructs to be measured. The framework should describe the curricular area domains to be assessed and the aspects of the educational context to be addressed.
- Depending on the scope of the curricular area domains to be assessed, a matrix-sampling data-collection design may be necessary, specifying how the overall assessment should be divided into manageable parts that can be assigned to individual respondents (usually students). For a study measuring trends over time, the design should specify characteristics of link items and how technical design features such as linking blocks or link item sets are incorporated in the assessment design.
- The study design should ensure that the sampling plan identifies the target populations of the study and proposes a sampling design to provide precise and economical estimates of population parameters. This should also include guidance on handling sampling and target population issues arising over time for trend or longitudinal designs. The sampling plan should specify the sampling techniques to be used and the required sampling precision.
- The data-collection instruments (achievement tests, questionnaires, etc.) to be developed or adapted should be specified, together with the scope of the data collection effort and the resources necessary. This should include plans for piloting or field-testing instruments as necessary.
- The design should specify the data collection methods to be used (computer-based, print-based, etc.), 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.
- The design should include analysis plans for both the field test and the main data collection, showing how field test data will be analyzed to inform instrument development and the main data collection. Analysis plans should describe the psychometric techniques used in constructing cognitive and context scales and indices, and the analytic methods to be applied to the data to address the study questions.

The study design should include a reporting plan giving an overview of how the study's results will be reported, and a data dissemination plan.

Standard for Choosing a Computer-Based Assessment System

Purpose:

To provide guidance for the selection of a computer-based assessment system.

Standard:

For a computer-based assessment, the assessment system should include all essential applications needed for item authoring, adaptation and translation, instrument design, instrument delivery, scoring, and a database to securely store data and provide data exports. The applications within the system should be configurable to meet the needs of the study and fully integrated such that data exchange across the applications is seamless and secure.

Guidelines:

The applications in the computer-based assessment system may be part of a single assessment application suite, provided by a single entity (such as IEA or a commercial software provider), or be a combination of applications from a range of providers.

- The item authoring application should have the capacity to create all achievement and questionnaire item formats to be used in the study. Ideally, international study center staff should be able to author and edit item content in the system. However, it may be necessary in some circumstances for software developers to create and edit bespoke complex items.
- The adaptation and translation application should enable adaptation and translation of item content into all languages of the study. The application should include workflow management features; the facility to show a record of the adaptation and translation history (both viewable by users within the interface and exportable for archiving); a facility for reviewing items and review comments to be added by item; and a preview feature to display the translated item content as it would appear to respondents. The application should also be compatible with any relevant standard import and export formats used by translation tools (such as XLIFF).
- The computer-based assessment system should include the facility for assessment and questionnaire instrument assembly, by allocating items to instruments. It should be able to manage adaptive and matrix-sampling instrument designs and other forms of complex assessment designs.
- The system must include an application to administer the instruments to respondents. The application must support online delivery (using the internet) and/or offline delivery (locally on devices or local networks), as required by the study.
- The system must include administration functions for users to, for example, create and manage user accounts, generate status reports associated with the different system components and monitor the status of data collection.
- Where human scoring of student responses is required, the system must include a scoring application that can accurately display student responses to scorers, support scorer training, and the evaluation of scoring quality both during the scoring activities and by supporting scoring designs that include multiple scores per response to facilitate methods that evaluate and utilize measures of scoring agreement.
- The system must be able to save all response data (including process data) to a centralized secure database that is accessible by the international study center and, which is accessible by the international study center, and have the facility to export data in formats specified by the international study center.
- The system must be compatible with existing IEA software systems relating to, for example, sampling and data management.

Standard for Ensuring Accessibility Support

Purpose:

To ensure that the study is as inclusive of all members of the study population as possible, including respondents requiring accessibility support to facilitate their participation in the study.

Standard:

The standardization of instruments and procedures is essential to ensure the reliability, validity, and comparability of ILSA data across countries (and within countries). Accessibility supports for participants must be managed in a way that is consonant with this overarching requirement for standardization. Accessibility should be considered across the whole study, including, but not limited to, population definitions and sampling, instrument development, field operations, analysis and reporting. The implementation of accessibility support for participants must be consistent with the relevant policies and practices in participating countries and clearly specified and documented. Decisions regarding the limits to accessibility options in the study should be included in this document.

Guidelines:

The study's initial planning should include details of how the requirements for standardization and accessibility will be balanced.

- When determining population definitions and sampling protocols, decisions about the inclusion or exclusion of participants based on their accessibility needs should be documented and justified.
- Instrument development should be conducted considering the principles of Universal Design of Assessment (Dolan et al., 2010, World Bank, 2023). Where possible, the instruments should be designed to be accessible to as broad a range of participants as possible and further amenable to accommodations consistent with the study plans.
- Field operations procedures should explicitly address accommodation needs with clearly stated study policies and protocols for determining eligibility for participants to make use of accommodations, and for documenting and keeping record of the implementation of accommodations across all participating countries.
- Analysis and reporting plans should specify how data collected from participants using accommodations are to be analyzed and reported.

Standard for Data Protection

Purpose:

To ensure that IEA Studies comply with data protection laws (e.g., the European Union's General Data Protection Regulation) that address study participants' right to protection of personal data.

Standard:

To meet this standard, IEA studies should ensure that personal data are:

- processed in a lawful, fair, and transparent manner.
- collected for a specified, explicit, and legitimate purpose.
- limited to what is necessary in relation to the purposes for which they are processed.
- not stored for longer than necessary to achieve the purpose of processing.
- processed in a manner that ensures its integrity and confidentiality.

Guidelines:

The following activities should be implemented:

- Participants in IEA Studies should be provided with information about the processing of their personal data. This ensures the transparency of IEA Studies. This can be achieved by providing each participant (and/or their parents) with a Data Protection Declaration (prepared by IEA and adapted by the National Research Coordinator). Furthermore, IEA should collect information from all participating countries about their data protection laws and requirements and comply with those laws and requirements.
- Contracts determining the roles and responsibilities of each party, and safeguards to the processing of personal data must be in place, such as:
 - A Joint Controller Agreement between IEA and national research coordinators.
 - Contracts with (sub)-processors/contractors on national level.
 - Standard Contractual Clauses for transfer of personal data outside of the European Economic Area.
- National centers should have in place a data breach plan to mitigate and to prevent the occurrence of data breaches, which includes the following steps:
 - Prevent - Detect - Evaluate - Mitigate - Communicate - Information provided to the Data Protection Authority (if applicable) - Information provided to data subjects (if applicable) - Keep register of personal data breaches.
- To enable participants in IEA's studies to exercise their data subject rights, national research coordinators are required to put in place procedures enabling data subjects' requests, to notify IEA of data subject requests and to assist in their fulfillment. Such a request has to be submitted directly to IEA or to the national research coordinator and might include the following:
 - Request to access their personal data.
 - Request to rectify their personal data.
 - Request to erase their personal data.
 - Request to restrict the processing of their personal data.
 - Request to object to the processing of their personal data.

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 is an integral part of the study design. It should ensure that national target population(s) are defined operationally in such a way that they can be represented accurately using probability sampling survey methodology and are comparable across education systems. The sampling design should provide a detailed plan for estimating the parameters of the target population(s) ensuring an acceptable level of accuracy for the key study outcomes for relevant reporting groups.

Guidelines:

- The development of a sampling plan begins with the specification of the international target population. This definition must ensure that national centers are able to identify the targeted units (mostly students) unambiguously so that they are comparable across countries and are amenable to effective probability sampling.
- For practical and budgetary reasons, it is usually reasonable to allow exclusions of eligible students from this population. Valid criteria and acceptable thresholds for such exclusions must be agreed-upon, controlled for, and documented. As they jeopardize comparability, exclusions should always be kept at a minimum.
- The sampling plan should specify the sampling techniques to be employed. All IEA studies should adopt probability sampling designs. These often involve complex features such as stratification, multiple sampling stages, clustering, and varying selection probabilities, which provide for cost-efficient and effective samples. The rationale and procedures for implementing these techniques should be specified. If samples for multiple studies are selected simultaneously, the overlap-control techniques used must follow established methods and be documented.
- The sampling plan should further specify the required sampling precision and minimum sample size for each country, considering expected design effects. Depending on the conditions, these may vary by participating education system. The sampling design should represent the best compromise within the range of tolerable sampling errors, costs, burden, and other considerations such as a national requirement for results for specific population subgroups. If multiple stage sampling is employed, sample sizes should be specified and justified for each selection stage.
- Samples selected with varying selection probabilities require weighting of the sampled units. The sampling weights constitute the inverse of the selection probabilities at the various selection stages. The sampling plan should demonstrate the envisaged weighting scheme.
- A key feature of high-quality sample surveys is the aim for high participation and/or response rates. The sampling plan should include measures to encourage countries to achieve this goal. Thresholds for these rates must be determined and made transparent together with consequences in cases of failure. Procedures for dealing with unit nonresponse must be developed, and the potential impact of nonresponse on population estimates must be described. Determining replacements for primary sampling units (usually schools) is one measure to make up for sampling loss. It is extremely important that the replacement process be under the close control of the international sample design team, as it can be easily subject to abuse. For this reason, the use of replacements for individuals (students or teachers) is generally not recommended. Another method used to address nonresponse is the computation of nonresponse adjustments during the weighting process, as well as the use of nonresponse bias analyses. Specifying the procedures for implementing and documenting such adjustments and analyses is another crucial part of each sampling plan.
- In case of failure to reach the participation rate thresholds, countries may opt to conduct a nonresponse bias analysis to evaluate the potential for bias due to nonresponse. Guidelines for conducting such analyses should be agreed in consultation with the corresponding international study center.
- Finally, the sampling plan should specify how sampling errors will be estimated. The method chosen must account for the complex sampling design implemented in the study.

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:

To collect the data necessary to address the study questions, the study should choose the method that is most accurate, efficient, and effective, while minimizing the burden on data providers (i.e., questionnaire respondents, achievement assessment takers, etc.) and making the least demand on available resources.

Guidelines:

- Data collection methods should capitalize as much as possible on the efficiencies offered by information and communication technology and its expansion in school systems worldwide. The data collection plan should consider the infrastructure available within countries to support computer-based delivery and, where necessary, include alternative technical solutions for data collection within countries.
- Online computer-based data collection can improve data quality and reduce data entry costs compared to paper-and-pencil print-based achievement tests and questionnaires, as respondent data are captured directly from the point of administration. This approach has the added advantage of capturing event and process data and time-stamped response history.
- Online data collection also can enhance respondent motivation and increase data accuracy and reliability. Maximizing response rates by keeping the burden on respondents to a minimum should be a priority.
- The data collection plan should specify the data providers for each instrument (e.g., students, teachers, school principals, parents, subject matter experts), when and how often data are to be collected, and the collection, processing, analysis, and reporting procedures to be used.

Standard for Developing a Quality Assurance Plan

Purpose:

To provide evidence that the instrument development and data collection procedures of the study are of sufficiently high quality to support authoritative international comparisons.

Standard:

The plan should include provision for independent monitoring of the quality of instrument adaptation and translation and data collection activities. The international study center should ensure that all operational documentation emphasizes quality control as an integral part of every data collection activity.

Guidelines:

- The plan requires development of procedures for checking and re-checking materials, monitoring data collection activities, and making explicit provision for resources for quality assurance.
- The plan should include procedures for the independent verification of the accuracy of adaptations and translations.
- Data collection activities such as test administration, which may be conducted by school personnel, should be independently monitored. This may involve site visits to a selected sub-sample of schools. Such independent monitoring should be conducted by the national center and, as far as funds permit, by the international study center (see also Standard for Implementing Data Collection Quality Control Procedures).
- The plan should include details of methods for evaluating the quality of data and identifying anomalies that may have resulted from issues in the data collection.
- Unauthorized data file modifications should be monitored.

Standard for Preparing an Analysis Plan

Purpose:

To provide a detailed plan that clearly shows how the study data will be analyzed to achieve the study objectives.

Standard:

The analysis plan should list each of the study's aims or research questions and provide details of the techniques to be applied to address them, together with a justification for each technique. The plan should enable national research coordinators and other reviewers to evaluate the intended analytic techniques and statistical procedures in the light of the study's aims, scope, and resources.

Guidelines:

- For each research question there should be detailed specifications of how the data will be analyzed and reported. The analysis plan should ensure that the analytic techniques are appropriate for the data to be collected and the study's design.
- The analysis plan should describe the types of simple descriptive analyses that produce accessible and easily communicated results before proceeding with more powerful multivariate analyses requiring advanced complex techniques that are more challenging to explain.
- Where appropriate, the analysis specifications may be supplemented by table shells or draft graphics that outline the statistics to be reported in the international reports. These help ensure that the analyses focus on the statistics necessary to address the main research questions.
- The plan should make explicit the kinds of inferences that may be made from the results of data analyses, and the limitations imposed by the design and analyses.

Standard for Selecting Psychometric and Statistical Analysis Procedures

Purpose:

To ensure optimal psychometric and analytic procedures are chosen to address the study's aims.

Standard:

Psychometric and statistical analysis procedures should be based on established best practice for large-scale data analyses in the social sciences. Psychometric procedures should be aligned with the data collection design, with the types of items in the achievement tests and the types of response variables collected in the context questionnaires. Analytic procedures should be appropriate for modelling and analyzing the study data and provide accurate and reliable results in a timely manner while ensuring robustness against violations of assumptions.

Guidelines:

- In psychometric analysis involving the construction and validation of measurement scales, project teams are encouraged to employ advanced psychometric methods, considering the assessment and data collection designs, the item types, and the need for linkage across test forms, countries, and over time. Choosing psychometric models appropriate for question type, (e.g., multiple-choice, constructed response, Likert type, etc.) may involve consultations between subject matter and psychometric experts to ensure that the methods used guarantee the validity and reliability of the analyses and can withstand scrutiny for violations of assumptions.
- Similarly, in statistical analysis, project teams must manage the complexity of cross-national data sets by applying appropriate statistical procedures to conduct analyses in a timely and accurate manner. It is essential to use methods that are robust to variations in curriculum, language, and cultural contexts and avoid potential violations of underlying mathematical assumptions. This approach guarantees that the analysis results reach accepted reliability, validity, and comparability standards across diverse education systems.
- Given the immense cost and efforts taken in IEA assessments, there is a need to choose the most appropriate advanced psychometric and statistical methods. At the same time, it is important to bear in mind that the results may have to be explained to a non-technical audience of subject-matter experts, researchers, and policy makers. A collaborative approach to developing a communication plan should be adopted to ensure that the methods chosen for the study are accepted and understood by stakeholders and users of the data.

Standards for Developing Data Collection Instruments

Section Overview

This set of standards applies to the development of instruments for measuring student achievement in school curricular areas, and for questionnaires to collect background, contextual, experiential and attitudinal information from students, parents, teachers, and school leaders. These standards are designed to ensure that IEA develops valid and reliable assessments and questionnaires through which knowledge, skills, abilities, attitudes, affective responses, and personal background characteristics are measured. The standards also ensure that the procedures followed, and the criteria adopted, are appropriate for the intended use of instruments.

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

- assessment frameworks defining the curricular domains and constructs to be assessed
- detailed specifications for the instruments addressing the assessment frameworks
- development of achievement and questionnaire items
- scoring guides for human scoring of achievement items
- verification of the translations of instruments in the different languages of the participating countries
- field testing the data collection instruments and procedures.

As noted in the previous section, the main parameters of instrument development are set by the study design, including the ages/grades of the target populations, when and how often the data are to be collected, and the assessment administration and data-processing procedures to be followed. The study design also should identify the achievement assessment and questionnaire instruments that need to be adapted or developed and outline the scope of this effort and the resources necessary. The instrument development process begins with the assessment frameworks, which define the curricular area dimensions, domains, and topics to be assessed, and the educational and environmental contexts to be addressed. Building on the assessment goals outlined in the frameworks, the assessment specifications provide the many details necessary to construct the achievement assessment and questionnaire instruments. These include:

- specifications about the item delivery mode(s), item and response formats, and item response data to be collected
- the timing associated with the instruments
- the weighting associated with the aspects of the framework
- the numbers and types of items
- guidelines on how to be sensitive to the different cultures represented by the participating countries.

The achievement and questionnaire items are developed in accordance with the content and psychometric specifications. Each achievement item should measure an important topic in the framework and be clear and accurate. Given the broad coverage of large-scale assessments, individual student's achievement assessment booklets (or computer-based booklet equivalents) typically contain only a portion of the achievement item pool. However, the item pool as a whole should provide coverage of the framework and be of appropriate difficulty. Questionnaires should focus on constructs prescribed by the framework and contain no more questions than are necessary to obtain that information in a valid and reliable way. However, both achievement and questionnaire instruments should have sufficient items from each design domain to provide reliable and valid measurement scales, where these have been planned to be constructed and reported.

Scoring guides for achievement items should be drafted by content experts in consultation with psychometric advisors as an integral part of the item development process and should include clear, unambiguous instructions to scorers about how to score every possible type of response. It is a hallmark of IEA studies that national centers actively contribute to the development and review of the assessment and questionnaire instrument. The international study center should ensure opportunities for this active contribution as part of each stage of instrument development.

Prior to field testing and again to the actual data collection, the instruments should be translated into their languages of instruction by the national centers. The international study center should verify these translations. These translations should be verified by the international study center to ensure that the meaning and difficulty level of the items have not changed from the international version.

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 main survey data collection by providing participants an opportunity to conduct all aspects of sampling, instrument preparation, data collection, and scoring using the procedures planned for use in the main survey. Field testing is particularly important for assessment items, as it provides valuable information about their psychometric properties in a realistic school setting, with students representative of the target population. Such information is predictive of how the items would perform in a full assessment and is useful in assembling final assessment forms that meet both content and psychometric specifications. Field testing is also important for reviewing questionnaire-based measures to test dimensionality in different contexts, the reliability of measurement scales, and to use the distribution of responses to evaluate the adequacy of response options. The outcomes of field trial operations should also be used to refine operational procedures for the main survey.

Standard for Developing Assessment Frameworks

Purpose:

To provide the conceptual and operational foundations for the development of the data collection instruments and methods for the study.

Standard:

The assessment framework should define the age/grade level of the student target population and the domains and constructs that will be measured and reported as outcomes in the study. These include cognitive domains assessed using achievement items (e.g., reading, mathematics, science, civic knowledge, computer and information literacy, computational thinking etc.), as well as constructs measuring attitudes, values behaviors and dispositions (e.g., self-efficacy, enjoyment of learning, civic engagement, trust in civic institutions, beliefs in the value of ICT use etc.).

The framework should also describe the aspects of respondents' personal backgrounds and contexts that are expected to be associated with the outcomes being measured. This may include an indication of the nature of any predicted relationships among data collected in the study with a view to framing the analysis and reporting of data.

Guidelines:

The assessment framework should include references to current scholarly literature and other relevant information sources to support the construct definitions, conceptual models, and instrument content and formats described in the framework. Suggestions for new content, comments and reviews by experts from the participating countries should be invited and considered when developing the framework. The framework should be published in an accurate and timely manner.

The assessment framework should describe the constructs and domains to be measured, and the procedures followed for defining them.

- For assessments of student achievement, the framework should:
 - Specify the dimensions, areas, topics, and subtopics within the domain that are age/grade-appropriate for the target population.
 - Describe what students are expected to know or be able to do in the domain being assessed.
 - Describe and provide examples of the instruments used to measure achievement.
- For measures of attitudes, values, behaviors, and dispositions, the framework should:
 - Define the constructs and explain, with evidence, their relevance to the study.
 - Describe and explain the content of the instruments to be used to measure attitudes and values.
- The assessment framework should describe the educational and environmental contexts (societal, home, school, classroom, etc.) about which data will be collected. This should include:
 - A description of the conceptual model that is being used to investigate the relationships among context measures, and between context measures and outcome measures.
 - A description of the contextual variables being measured in the study, and how they are being measured.
- The framework should, where relevant, consider the relationship between data collected using the instruments for any given administration cycle and those of previous and possible future cycles.
- Where applicable, the assessment framework should provide an explanation of how process data (data collected in a computer-based system that records respondents' interactions with the instruments as they complete them on a computer) are to be used. This should include:
 - A rationale for the collection of process data; and
 - A description of the process data to be collected.
- The assessment framework should provide guidance about how the data collected in the study can be reported.

Standard for Developing Specifications for Achievement and Questionnaire Instruments

Purpose:

To ensure that the data collection instruments will reflect the domains, constructs and concepts explicated in the assessment framework.

Standard:

The specifications for the achievement and questionnaire instruments should describe the content, format, and delivery method(s) of each instrument; 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 instruments are administered. They also should describe how responses are to be scored. The specifications should include details of the instrument design, including the allocation of items and questions to assessment and questionnaire forms, as well as plans for the use of content to link data to previous and possible future study cycles. The specifications should include details of any process data to be collected.

Guidelines:

The achievement and questionnaire instruments should be developed with explicit reference to the conceptual models and constructs outlined in the assessment framework. The degree to which the instruments reflect the specifications in the assessment framework should be used as one evaluation criterion during each step of the development and review process of the instruments.

The instrument development specifications should include:

- The delivery mode(s) and time allocated for respondents for the achievement assessments and questionnaires.
- The item and response formats to be included in the instruments.
- The proportions of the different item and response formats comprising the instruments.
- The design and layout of the instruments.
- Where relevant, the types of process data to be collected for each instrument and item type.
- The scoring methods for the various item and response formats.
- The allocation of items and score points to each assessment domain described in the assessment framework.
- The inclusion of trend items (where relevant) in the instruments, including the number of trend items.
- The number of items to be included in the field trial instruments and, based on defensible assumptions of about item loss following the field test, the number of items to be included in the main survey instrument.
- Plans for linking data from the current study cycle to possible future cycles; and
- An assessment design that specifies the allocation of items to test and questionnaire forms.

Standard for Achievement and Questionnaire Item Development

Purpose:

To ensure that the achievement and questionnaire instruments reflect the assessment framework specifications and provide fair, reliable, and valid measures of the domains, constructs, and concepts explicated in the framework.

Standard:

All aspects of the data collection instruments should be clear, accurate, and easy to understand, including the directions, stimulus materials, and items. The assessment and questionnaire items should accurately represent the domains, constructs, and concepts explicated in the assessment framework, and be appropriate for the population of respondents. The items should be engaging to respondents, be free of personal, social, cultural, or language bias, and be scorable reliably and efficiently. The item development process may include small-scale pilot studies to refine the item writing process and should include a full-scale field trial before the main data collection. The procedures used to develop the items should be clearly documented.

Guidelines:

Item development is an iterative process of creation, review, and revision. At different stages of the process, opportunities should be provided for content and measurement experts, and study stakeholders to review the items and associated content (such as stimulus material and scoring guides). The contribution of participating countries to the development process is essential in minimizing the risks associated with linguistic and cultural biases affecting the validity and reliability of the items. Where feasible, item development should be informed by information gathered from small-scale pilots of items.

When items are being developed, reviewed, and revised, each item should be evaluated, as a minimum, from the following perspectives:

- Content validity
 - Does the item material meet the instrument specifications?
 - Does the item material clearly represent relevant domain, constructs, or concepts?
 - Would the item stand up to broader stakeholder and public scrutiny?
- Clarity and context
 - Is the item coherent, unambiguous, and clear?
 - Is the item worthwhile, and relevant?
 - Does the item assume prior knowledge and, if so, was this assumed to be acceptable or part of what the instrument is intended to measure?
 - Is the reading load as low as possible?
 - Can the item be translated without impediment (such as through idiom or language-specific structures)?
 - Does the item repeat content already addressed in one or more other items and, if so, is this appropriate?
- Format
 - Is the proposed format of the item the most suitable for its content?
 - Could the item format create any construct-irrelevant challenges to respondents?
 - Is the item format likely to demotivate respondents?
- Respondents
 - Does the item sit within the reasonably expected range of ability, age, and maturity of the respondent population?
 - Is it clear to respondents how their responses should be completed?

- Is the item cross-culturally appropriate, relevant, and sensitive?
- Is the item free of construct-irrelevant bias against known subgroups in the target population?
- Are respondents' responses to an item likely to be affected by their experience of, or exposure to, the contents of other items in the instrument and, if so, is this appropriate?
- Scoring
 - For multiple-choice achievement items, is the key (the correct answer) irrefutably correct?
 - For multiple-choice achievement items, are the distractors (the incorrect options) plausible but also irrefutably incorrect?
 - Are the proposed scoring rules for an item consistent with the underlying domain, construct, or concept being measured?
 - Are the proposed scoring rules for an item (including a scoring guide where relevant) clearly expressed and unambiguous?
 - Do the scoring rules account for the full possible range of student responses?
- Process data
 - Are any proposed process data to be collected when the item is being completed relevant to the item content and completion process?
 - Where process data are intended to contribute to scoring of the item, is the planned process data appropriate for this purpose and does the proposed interpretation of these data match the conceptual content of the item?

A field trial should be conducted in which the items are administered to a sufficiently large sample of respondents, representative of the target population(s), to allow for an analysis of the measurement properties of the items. The results of the field trial analyses should inform decisions about the items to be included in the main data collection.

Any item revisions between the field test and main survey should be kept to a minimum and avoided where possible. Such changes should only rarely be made if evidence is overwhelming that a clear error in the item is being corrected, and the study center is confident that the change will correct a known problem and will not introduce any new problems.

Metadata about the properties of all items should be recorded during item development. The metadata record (codebook) should include details of all relevant item and variable properties that are required for use in the data management, cleaning and analyses as well as any additional properties that may be useful for item review and dissemination.

Any process data to be collected should also be described, defined and clearly documented with the item information.

The item development process should be clearly documented.

Standard for Developing Scoring Guides for Human Scoring

Purpose:

To ensure that the scoring guides provide complete and unambiguous instructions of how responses are to be scored by human scorers.

Standard:

Scoring guides for achievement items should include clear, unambiguous instructions to scorers about how to score every possible type of response. The scoring guides should include examples of genuine responses, that illustrate the key characteristics of different score categories within each item. Example student responses should be provided as training materials that countries can use to train scoring staff.

Guidelines:

- Where specific categories of responses to an item are to be assigned specific scores, the categories should be mutually exclusive (i.e., without overlap) and exhaustive (i.e., as a set, they should account for all possible responses). The scoring logic to be applied to each item should be explicit and clearly described. Each score category should be illustrated by example responses. These examples should, where feasible, be selected to:
 - Clearly reflect the defining characteristics of the category.
 - Represent borderline responses between categories; and
 - Demonstrate unusual, ambiguous, or difficult to interpret responses.
- The choice of, and number of example responses can be informed by the challenges associated with scoring the item. Ideally, the responses should be obtained from a data collection exercise with students in the same grade as the target population (such as from a pilot or field trial data collection). Where necessary and appropriate, example responses can be generated by the international study center team.
- In addition to example responses included in the scoring guides themselves, it is advisable to prepare scoring training materials that can be shared with national centers to train their scoring staff. These training responses should be drawn from the field test or even the main data collection. A range of student responses should be selected for each item. These training materials may include 'borderline' responses that are near the boundaries of adjacent score categories for an item and should be presented with an explanation of how such responses are to be interpreted and scored.
- Prior to the field trial, and again before the main data collection, the international study center should conduct international training in item scoring for representatives from participating countries. These "training the trainers" sessions can provide models that national research coordinators can replicate in their home countries, so that scoring staff can develop a shared understanding of how the scoring guides can be validly and reliably implemented in each country.
- Scoring guides should be continually revised throughout the item development process. Specifically, field-test versions of the scoring guides should be refined and finalized in response to feedback received during the field-test scoring training meeting. The international study center should further seek feedback from national centers about the field test scoring. This should collect information about specific items or response categories within items that were difficult to score, and any suggestions for improvements to the scoring guides and support materials. This information should contribute to revision of the scoring guides between the field test and the main survey. In addition, student responses collected during the field test should be used to supplement the existing sets of example and training responses. The scoring guides for the main survey should be further refined and finalized in response to feedback received during the main survey scoring training.

Standard for Translations and for Verifying Translations

Purpose:

To ensure that the achievement and questionnaire items are translated from the international versions into the target languages without changes in meaning or difficulty to provide comparable data across countries, cultures, and assessment cycles.

Standard:

When translating achievement items or modifying them for cultural adaptation, participating countries should ensure that the following characteristics remain the same as in the international version: the meaning of the item; the reading level of the text; the difficulty of the item; and the likelihood of another possible correct answer for the test item. The meaning and content must remain the same for questionnaire items when translating or adapting to cultural contexts. There should be an independent international verification of all assessment and questionnaire instruments.

Guidelines:

- If possible, national research coordinators should use a system for translating the data collection instruments that incorporates an independent check on the translation. For example, the national center could produce two independent translations and then have a third party reconcile the differences between them.
- In general, the translators' work includes the following:
 - Identifying and minimizing cultural differences.
 - Finding equivalent words and phrases.
 - Ensuring that the reading level is the same in the target language as in the international version.
 - Ensuring that the essential meaning of the items does not change.
 - Being aware of changes in layout due to translation.
- Online translation systems as well as AI supported translation systems can be very helpful in producing translated text. However, national centers should carefully monitor the output of such systems to ensure the quality of the resulting translations.
- Where translations are subcontracted to an external agency, the national center should retain responsibility for the final product and for ensuring high quality before submitting translations to IEA for verification.
- The international study center should explain 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. The survey operations documentation should include materials for the participating countries that describe the translation and cultural adaptation procedures and the process for translation verification. Countries document any deviations in vocabulary, meaning, or item layout from the international versions.
- International translation verification is conducted by IEA in conjunction with an independent translation company. As draft translated instruments are received from the participating countries, they should be verified by professional translators by comparing each translated instrument with the international version and documenting any differences. Any deviations from the international instruments that the translation verifier judges as affecting the results should be addressed by the national center. Examples include incorrect ordering of response options in a multiple-choice item, mislabeling of a graph that is essential to a solution, or an incorrect translation of an assessment question that renders it no longer answerable or suggests the answer to the question. IEA should verify that all deviations judged to affect the results have been rectified.
- Following the field test, the international study center should conduct a thorough review of diagnostic item statistics to detect any translation errors that were not corrected to minimize the possibility of translation errors in final versions of the test instruments.

Standard for Field Testing Data Collection Instruments and Procedures

Purpose:

To provide evidence of the suitability of the achievement assessments and questionnaires, and data collection procedures for the main study data collection.

Standard:

Prior to the main data collection, all achievement and questionnaire items, tasks and directions should be field tested on a population as similar as possible to the target population. Field test achievement and questionnaire instruments should be as close as possible to those planned for the data collection in the main survey, as should the data collection procedures and systems, including computer-based delivery systems. Sample sizes should be large enough to provide meaningful diagnostic statistics about the behavior of the field-test instruments. 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:

- The field test instruments should be administered in each participating country to a sample from a population similar to the national target population. Ideally, the field test school sample should be drawn in a prior year, but at the same time within the school year, and using the same probability-sampling methods as the main data collection sample. This ensures that the field test sample cannot include respondents who will provide data in the main data collection.
- The field test samples should be large enough to provide useful diagnostic statistics for item selection and refinement. As an example, the TIMSS and PIRLS field tests typically involve 25–30 schools and provide at least 200 student observations for each cognitive item.
- The field test should include enough items to permit assembly of the final data collection instruments from the field-tested items. A good guiding principle is that about twice as many assessment items as will be needed should be field tested.
- In situations where new domains or item formats are being used, or in countries with limited experience of conducting field operations, it may be useful to conduct some small-scale pilot testing before the field test.
- The field-test data collection procedures should, as much as is possible, be the same as those planned for the main data collection. This will ensure that assessment and questionnaire item characteristics and operations procedures are realistically evaluated in all participating countries. This also gives the opportunity to evaluate the procedures and, where appropriate, to refine them for the main survey.
- Timing for the field-test forms should correspond to the timing planned for the final assessment. This process will provide better information about the length of the assessment forms and questionnaires relative to the time assigned to complete them. For online data collection, response time information provided by the assessment system can also be useful for this purpose.
- Field-test instruments should be reviewed by content experts, assessment development experts, editorial specialists, and proofreaders (see also Standard for Achievement and Questionnaire Item Development).
- Field-test assessment item responses should be scored as planned for the main survey data collection. 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. Information from the field trial should be used to review and, where appropriate, refine the main survey scoring guides and procedures.
- Psychometric item analyses based on field-test responses should be conducted. These analyses should include the psychometric techniques planned for the main data collection and should provide item difficulty and discrimination estimates, statistics about differential item performance by country and by gender, and data about the reliability of the scoring procedures. This information can 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 assessment.

- The measurement properties of the questionnaire items in the field test should also be examined. Item response distributions and scale properties should be evaluated to inform data collection for the main survey. Low response rates may indicate 'speediness' or lack of interest by the respondents. For sets of items intended to represent constructs, the psychometric techniques planned for the main data collection to investigate the validity and reliability of the resulting scales should be applied.

Standards for Data Collection and Processing

Section Overview

This set of standards applies to the collection and processing of assessment data.

In designing data collection systems, procedures, and materials, it is important to remember that IEA assessments are crucially dependent on school systems, staff, and students to provide the necessary data. Because student assessments disrupt school schedules and impact the amount of time that can be used for instructional purposes, the burden imposed by a proposed assessment should be carefully evaluated, with due consideration given to the fact that providing assessment data is generally a voluntary activity and not a primary responsibility of the school system.

The plans for data collection should focus on international requirements for standardization, while recognizing national variations in how school systems are structured and operate. The data collection design should be implemented through carefully defined practical procedures, capitalizing on the efficiencies afforded by information technology where possible. Data collection procedures should be fully documented, in manuals and software systems as appropriate, and national center staff trained to conduct the procedures correctly. Proposed adaptations to procedures should be reviewed by the international study center for international comparability and for conformity with the study intent and design.

Quality control during data collection should be built into the study plans and should include internal mechanisms to ensure procedures are documented and implemented correctly and external reviews conducted by independent monitors. The data collection documentation for each country should describe all aspects of the data collection, including all adaptations to, and deviations from, the international procedures. Procedures for processing the data must also be developed and documented. These may include procedures for scoring student responses, both human scoring and scoring supported by artificial intelligence. The goal is to ensure that each country's data are prepared in the same way and subjected to the same consistency and quality checks. Procedures for reviewing and checking source documents should also be specified.

The following sections define standards for each of the major components of data collection and processing. The standards generally apply to activities at the international level. However, given that data collection is the responsibility of the participating countries, standards also have been provided for activities at the national level 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(s) with minimal risk of bias, and that the sampling procedure will accommodate the computation of sampling weights and variances.

Standard:

Samples for IEA studies should conform to the study sampling design and use methods derived from sound and defensible sampling theory. Study samples should meet specified quality standards regarding coverage and sample size as specified in the sampling plan.

Guidelines:

- The international study center should provide a sampling manual that specifies each step of the sampling process. The international study center should also provide training and support for national centers to provide the necessary information to complete the sampling procedures.
- The samples must be verifiable, i.e., it should be possible, with readily available documentation, to replicate any selected sample. The documentation should include: a written record of the sampling method and process; a detailed description of the sample design parameters; descriptive information essential in determining the quality of the resulting samples (such as the extent of coverage, exclusion rates, and participation rates); and data from which to derive the sampling weights for the computation of all survey statistics.
- Target populations in IEA studies are usually hierarchical, for example, students within grade levels within schools or in classrooms within schools. Multi-stage sampling designs are suited well to this situation and are used extensively in IEA studies. Often, the national population is defined in terms of a list of schools containing eligible students. Lists of schools usually serve as the sampling frame for the first stage of the sampling design, with schools being the primary sampling units (PSUs). Lists of classes and/or students serve for the subsequent sampling stages within PSUs. These sampling frames are provided by national centers and must be thoroughly checked against officially available statistics. In particular, the following problems must be avoided:
 - Under-coverage: eligible units of the population (e.g., schools, classes, or students) are not included in the sampling frame. Such units would have no chance to be selected for the survey, and survey results would not apply anymore to the whole target population.
 - Over-coverage: 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 ultimately have to be discarded. The result would be a smaller sample size than expected.
 - Duplication: 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 contains incorrect or outdated information. Such errors can lead to coverage problems, errors in stratification and inappropriate selection probabilities. Where errors are detected, they should be corrected.
- Stratification of the sampling frame prior to sampling should be considered 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 minimize sampling error by eliminating between-stratum variance.
 - Explicit stratification (where a separate sample is drawn within each stratum) should be considered whenever disproportionate sampling allocations are desired. This could be necessary if, for example, oversampling of small subpopulations relative to larger ones is needed to support sufficient statistical power for group comparisons.
 - Implicit stratification should be considered 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. Regardless of the stratification methods used, strata should be large enough to contain several sampled units. Two sampled units per stratum are the fewest that can yield within-stratum sampling errors.

- Replacement schools should be identified when the school sample is drawn. Each replacement school should be linked to one and only one sampled school. Replacement schools should be adjacent to the sampled school in the school sampling frame. In each stratum, schools should be listed in a systematic order so that the adjacent school is similar to the sampled school.
- Wherever possible, school sampling should be done centrally by a small group of sampling experts, and in a standardized way. For within-school sampling conducted at national centers, sampling software should be provided by the international study center to assist with sampling activities. Sampling software should not only reduce the burden on national centers but also ensure standardized procedures and produce the tracking information that will be needed for computing sampling weights and for auditing purposes.
- All regulations on data confidentiality that apply in the participating countries must be obeyed. All data should be exchanged and stored in a secure digital environment. Regulations regarding the protection of personal data such as, for example, the General Data Protection Regulation (GDPR) of the EU should be enforced and considered in all sampling related procedures.

Standard for Planning for Data Collection

Purpose:

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

Standard:

Planning for data collection involves implementing the international study design in the form of standardized 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 study center and the national research coordinators 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 study center should:

- Involve representatives of the participating countries in deciding the data collection plan and schedule of data collection.
- Develop a detailed data collection plan that includes the following:
 - Descriptions of the target populations and rules for including/ excluding population members.
 - A clear statement that any special preparation of students in the selected schools before the main data collection is not admissible as it would invalidate the representativity of the results.
 - Procedures for selecting school and student samples and identifying other respondents such as teachers and school principals.
 - For computer-based assessments, instructions for selecting, implementing, and monitoring the data collection environment(s).
 - For computer-based assessments, instructions and test software to assess the readiness of schools for the computer-based data collection.
 - For computer-based assessments, instructions for all users to operate the computer-based data collection system correctly.
 - 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.
 - A reminder that national data protection guidelines and other legal requirements for conducting assessments in schools must be ensured.
 - Response rate requirements and procedures for using replacement respondents.
 - Requirements for data collection staff organization and training.
- Use the plan to develop manuals specifying the procedures, software, and materials appropriate for each level of data collection (i.e., national research coordinators, school coordinator, test administrator, etc.). Here, it will be necessary to:
 - Give representatives of the participating countries multiple opportunities to review draft plans, manuals, software, materials and forms.
 - Conduct field test(s) of procedures, software, forms and materials, and make appropriate modifications.

- Develop a training program for national study staff that includes data protection training and ensuring the sample's representativeness.
- Develop a program for monitoring the quality of the data collection.

The national research coordinators should:

- Carefully review all study procedures, software, and materials for compatibility with national school system characteristics.
- Follow specified procedures for implementing necessary modifications in procedures and materials.
- Adapt international materials, as necessary, and obtain international approval.
- Prepare all administrative manuals, materials, software, and forms.
- Develop informational materials and a plan for securing permission and support from the requisite administrative levels.
- Make sure all data protection guidelines and other legal requirements are met.
- Notify participants as early as possible.
- Develop a plan for selecting, hiring, and training data collection staff.
- Develop a plan for administration and monitoring of data collection activities.

Standard for Minimizing Response Burden

Purpose:

To ensure that data collection activities are minimally intrusive and burdensome, so that non-response is kept to a minimum.

Standard:

Since burden and non-response are related to each other, all data collection procedures and materials should be carefully reviewed to prevent unnecessary burden, recognizing that participation is generally voluntary and not a primary activity for respondents.

Guidelines:

- As materials and procedures are developed to implement the design, they should be reviewed for clarity, ease of use, and simplicity. All instructions for respondents should be clear and easy to follow.
- Realistic time limits should be established for all administered measurement instruments, keeping in mind who the respondents are.
- Instruments should be constructed to be comfortably completed by respondents within their specified time limits. National centers should be given clear guidance about the limits to increased burden by adding content to national extensions to measurement instruments.
- Data collection procedures should be as simple as possible and clearly described. For computer-based data collection, the procedures for accessing and logging into the system should be as simple as possible.
- Students should be provided with practice questions to familiarize them with the item formats, and, for computer-based assessment, the test interface used in the assessment.
- Data collection should be scheduled at the convenience of the respondents, to the extent possible, recognizing that schools are very busy places and that participation in the study is often voluntary.
- Use the results of field tests to check that estimates of burden are realistic and acceptable.

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 evaluate each trainee's progress as they work through the training program.

Guidelines:

Staff from the international study center should complete the following tasks:

- Develop manuals that clearly describe all data collection procedures and materials and are clearly written so that national center staff will be able to use them with minimal training.
- Determine whether the data collection plan requires staff to have any special training or skills. If necessary, the study center should develop a training program that includes these skills and ensure that national staff receive this training.
- Countries may vary in their approach to data collection. For example, some countries use school staff to collect data and administer assessments, other countries use national center staff, and others hire data collectors specifically for the study. Regardless of the approach taken, it is important that all data collection staff receive the necessary skills and training.
- For computer-based assessments make sure the assessment application is available in advance, so that the countries can demonstrate the application and provide hands-on training to test administrators.
- Evaluate the procedures used in the field trial data collection and use this information to modify procedures for the main data collection as required.

Staff from national study centers should complete the following tasks:

- Adhere to the international data collection guidelines.
- Depending on the level and experience of data collection staff, adapt the international manuals, materials and training plans to cover points not covered or experience that may be lacking.
- Provide hands-on data collection training. Data collection staff should have the opportunity to practice all procedures (e.g., administer the assessment, reading out the instructions, fill out the forms, transcribe the records before they have to use them).
- In computer-based assessments, test administrators should practice the log-in procedure and, depending on the assessment-mode, data uploading. It is also recommended that test administrators are familiar with the items used in the testing platform and applications used in the testing session.
- Ensure that the data collection training includes troubleshooting and problem-solving procedures to be used if problems occur during the data collection sessions.
- Hire and train sufficient backup or substitute staff. If relying on local school staff as data collectors, ensure that backup staff are available within the schools.
- For computer-based assessment, it should be avoided that heavy equipment may need to be carried to the schools and that the test administrators have the capacity (including support) to handle this.
- Provide a contact person who is knowledgeable about the study and will be available to answer questions from data collection staff during the data collection period, especially during the testing time in the schools.
- Monitor field activities and identify problems encountered so that corrective action can be taken quickly.

Standard for Implementing Data Collection Quality Control Procedures

Purpose:

To provide evidence that data collection is conducted in accordance with the established study procedures.

Standard:

Quality control during data collection should be integral to the study at the national and international levels. Quality control includes internal mechanisms built into each stage of data collection to ensure procedures are implemented correctly, and external reviews conducted by independent monitors.

Guidelines:

- Quality control should be incorporated into study procedures, forms, and materials as they are being developed and explained clearly in study documentation such as manuals. In addition, it is essential to monitor data collection activities at the national level both internally (i.e., by the national study center) and externally (i.e., by the international study center).
- The international study center should employ independent International Quality Control Monitors to observe the data collection activities in a subset of schools (for example ten percent), record key information, and report that information back to national and international study centers. National centers are encouraged to supplement this effort with a program of visits by National Quality Observers.
- If employment of independent monitors is prohibitively expensive or not possible for other logistical reasons, alternative procedures should be sought. For example, reports from test administrators can be compiled and analyzed.
- Quality control procedures implemented during the field trial should be used to identify and correct quality problems prior to the main data collection. Quality control activities should be clearly documented and reported either to the international study center, in the case of national quality control, or in technical documentation, in the case of international quality control.

Standard for Documenting National Data Collection

Purpose:

To ensure that the national data collection activities can be reviewed by the international study center and documented for use by later users of the data.

Standard:

The data collection documentation for each country should describe all aspects of the data collection, including all adaptations to and deviations from the international procedures.

Guidelines:

- National study centers should fully document national data collection activities, including problems encountered and solutions applied. All agreed study procedures should be followed unless agreed otherwise with the international study center.
- The international study center should specify procedures for countries to document deviations from the source measurement instruments for each language version of the instruments to be administered in each country. These should clearly define permissible modifications to the instruments and provide instructions on how deviations are to be documented if not recorded automatically during the translation of the instruments.
- Terms in the source measurement instruments where cultural differences are expected (e.g., levels of education) should be considered as mandatory adaptations, and countries should be required to document the terms used in their instruments. The international study center should provide guidelines for permissible adaptations.
- National centers should collect information from staff directly involved in data collection activities (e.g., test administrators) to document all aspects of the data collection.

Standard for Planning Data Requirements, Data Saving, and Data Processing

Purpose:

To collect all data necessary to achieve the study objectives completely, consistently, and economically and to process them in such a way that the workflow is uninterrupted and coherent.

Standard:

The international study center should plan all data requirements and data saving when setting up the assessment system. The planning should relate to all data that is to be collected during the assessment. The planning of data processing should include all steps in the process chain.

Guidelines:

The planning of data requirements and general data saving should include all assessment and questionnaire responses as well as relevant events, actions, and processes associated with the responses.

The planning should consider which general and specific attributes should be collected for the responses. Attributes are the defined events and actions associated with responses such as timestamps, screen identifiers, item names, etc.

The planning should also specify the data and file formats that are made available by the assessment system. The data processing planning should include data retrieval from the assessment system, machine scoring, and human scoring including interface definitions and data transfer protocols.

In addition, the data processing planning must consider all data integrations, checks for integrity and consistency within and between data sets. The data processing system to be used should be selected with consideration of the characteristics of the data, the complexity of the data to be collected and the resources available (e.g., available infrastructure and economic resources).

Finally, the planning should consider the documentation effort to be made and how the data should be disseminated.

Standard for Planning for Data Preparation and Processing in Paper-based Assessments

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 study center that conform to the international data structure, (ii) are comparable internationally, and (iii) are free from error as far as possible.

Standard:

National coordinators should plan their data preparation activities in good time and ensure that they have all the information and understand the procedures necessary to create data files that conform to the international data structures.

The international study center 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:

The national coordinators should ensure that they fully understand all data preparation and processing procedures. The international study center should provide participants with manuals describing all aspects of data entry, including the assignment of identification numbers, data formats and file structures, and should provide training and assistance to national coordinators.

Standard for Human Scoring of Student Responses

Purpose:

To ensure that human scoring of student responses provides valid and reliable data with respect to the domains, constructs, or concepts being measured by the items being scored.

Standard:

The international study center should provide representatives from participating countries with training so that they can conduct scorer training and oversee scoring operations in their own countries. In turn, countries should replicate this training with their own staff to ensure high quality scoring of all assessment items. Scoring accuracy and reliability should be monitored and recorded. Ideally, any given response would receive the same score regardless of the person scoring the response, the country in which the scoring is taking place and the language of the response.

Guidelines:

Scorers participating in the international training and those running national operations should be qualified to complete the work allocated to them. Where a curricular area is being assessed, the scorers should have qualifications in that area, and, ideally, experience in teaching in that area at the grade level being assessed. Scoring staff should be trained to score every item that they will score.

The scoring training should include, as a minimum:

- An introduction to the study, including its purpose, target population and a summary overview of its procedures.
- Information about the learning domain/achievement construct to be scored with reference to the Assessment Framework and, where relevant, the existing achievement scale(s) derived for the study.
- Explanations of the scoring procedures, including training on the use of any computer-based scoring systems being used.
- A detailed explanation of the contents of the scoring guide for each item.
- An opportunity for scorers to score a range of 'training' responses for each item. The training responses used should:
 - Be selected from a set of genuine responses (or where this is not possible, contrived responses) from students who are representative of the target population.
 - Clearly reflect defining characteristics of each score category for the item.
 - Represent the characteristics at the borders between adjacent score categories; and
 - Are unusual, ambiguous, or difficult to interpret.
- An opportunity for scorers to raise questions of clarification and interpretation regarding the scoring guides and the training responses. Ideally the training for a given item will conclude only when scorers have indicated that all their queries have been adequately addressed.

To collect data on the reliability of the scoring effort, the international study center should provide countries with a procedure for organizing scoring so that a random sample of responses is blind scored by a second scorer.

While countries may choose to subcontract scoring of assessment items to an external agency, the national center remains responsible for the accuracy and reliability of the scoring activity. The national center should ensure that the external agency adheres strictly to all scoring guides and procedures.

Standard for Machine Scoring of Student Responses

Purpose:

To ensure that the use of machine learning and artificial intelligence to score student responses provides valid and reliable data with respect to the domains, constructs, or concepts being measured by the items being scored.

Standard:

Artificial intelligence technologies and methods are evolving rapidly and their potential to score responses should be considered as part of the study design and operations planning. This may include both the use of artificial intelligence scoring to complement human judgment and automatic scoring of responses without human involvement.

Artificial intelligence models used in scoring should be comprehensively trained. The accuracy and reliability of artificial intelligence scoring should be monitored and documented. Ideally, any given response would receive the same score from the artificial intelligence model used in the study, regardless of when the response was scored, or the origin and language of the response. The critical logic of the artificial intelligence models being used in scoring should be documented.

Guidelines:

- While human scoring of open-ended responses is likely to be necessary for some time to come, every opportunity should be taken to avail of the potential for machine learning and artificial intelligence to improve scoring reliability and efficiency. However, given the current state of the art, there is a need for careful evaluation of these approaches. These tools can be viewed with caution when planning to adopt automated scoring. Until it is shown that they meet or exceed the quality of human scoring it should be recognized that these technologies are tools to complement, rather than replace human judgment.
- The use of artificial intelligence in scoring should be specified in the study plan and communicated to stakeholders.
- The procedures used for developing the artificial intelligence scoring models should be fully documented, and the data used to train and evaluate the models should be stored and accessible for quality assurance and additional research.
- The configuration of the models to score each item should be clearly described, specified, and replicable.
- The reliability and accuracy of scoring should be monitored and evaluated. The methods of evaluation should be clearly documented and replicable. Where machine scoring is used to complement human scoring, the procedures and the relationship between the machine scores and human scores should be evaluated and described.
- Where feasible and relevant, artificial intelligence scoring documentation should include two layers: a technical explanation sufficient to support replication of the methods and a plain language explanation for a non-expert audience understanding the basic principles that underpin the artificial intelligence scoring and quality assurance processes.

Standard for Processing and Checking Data

Purpose:

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

Standard:

The data provided by national centers should conform to the international data structure and to the data definitions specified by the international study center for data saving. Data should also be free of internal inconsistencies. The international study center should ensure that all data files provided by participating countries comply with international formats and are free from error.

Guidelines:

- As the party ultimately responsible for the quality of their national data, national research coordinators should ensure that their staff and any external contractors follow all prescribed procedures for data preparation and implement all prescribed checks for data integrity and quality control.
- The international study center should encourage and support the national coordinators and their data suppliers in their data preparation and checking endeavors and ensure that data capture software includes all possible consistency checks to facilitate error identification and resolution at the point of entry.
- The international study center should check the data sets provided by each country against the international data structure and report all deviations to the providers. It also should report all extreme or unlikely values to the national centers and data suppliers for investigation and resolution.
- As an important step in ensuring high quality, internationally comparable data, the international study center should enable each country to conduct a data review by providing preliminary national level item analyses and univariate statistics.

Standard for Documenting Data-Processing Activities

Purpose:

To ensure that all modifications applied by the international study center to participants' data are recorded, whether such modifications were the result of structural reorganization or the result of errors.

Standard:

All errors and inconsistencies encountered in the data should be documented and communicated to national research coordinators. All modifications to data arising from the resolution of such issues should be 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:

The international study center should send national centers documentation of data-checking and editing activities of their national data that includes:

- A list of all data checking and editing procedures applied to the data.
- A record of all transformations applied to the original raw data.
- A list of all deviations from the international standard found in the data and instruments.
- A list of specific edits that have been applied to individual observations.
- A list of all edits that have been applied globally to the data files.
- A final database
- A codebook holding the metadata which describes the final database.
- Computational definitions for any variables re-coded or derived from other variables.
- Item analysis and preliminary univariate statistics for each variable.

When all edits and modifications to the data have been finalized, an international database should be produced. It is the responsibility of the international study center to produce documentation for users of this database that includes:

- A list of all data-checking and editing rules applied to the data.
- A codebook with metadata for each data file giving all the information necessary to access and analyze the data.
- An explanation of the structure of the database.
- A report that lists all adaptations and modifications made by national centers to the international instruments or procedures.
- Definitions of all variables re-coded or derived from other variables.

Standards for Analyzing Data and Reporting Results

Section Overview

Given the cost and effort required to develop and conduct an IEA study, it is essential that the data are analyzed and reported in such a way that the study aims are achieved and the results widely disseminated. Data analyses should be conducted using the most appropriate methods as specified in the analysis plan while adhering to established best practices. Once the plan has been reviewed and accepted it should be implemented faithfully.

Because IEA studies depend heavily on sample survey methods, it is important that appropriate sampling weights are computed and applied so that sample statistics are accurate estimates of population parameters. The standard for developing sampling weights provides guidelines for computing these weights and suggests procedures for checking their quality. While sampling weights are intended to ensure that sample statistics are consistent estimates of population values, any statistical estimation involves some error due to the sampling process, due to measurement uncertainty, and due to deviations from standardized data collection procedures. 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 potential bias that may result from non-sampling errors, including deficiencies in population coverage or low participation rates.

Student assessment data and questionnaire responses from parents, students, teachers, and school principals are generally reported in the form of measurement scales, which are expected to provide more reliable and valid information about the topics under study than responses to individual questions. For example, students' responses to a set of mathematics items are typically combined to provide an overall mathematics score summarizing the students' mathematics achievement conveniently and reliably. Similarly, a teacher's responses to questions about instructional practices may be combined to form a scale that indexes teaching style. The standard for validating measurement scales addresses the need to provide evidence that such scales are reliable and valid indices of the constructs they are designed to measure.

IEA studies are valuable only to the extent that their results are reliable, valid, and comparable across countries. It is important to promote wide dissemination and broad understanding of the results. The standard for presenting study findings stresses the need for 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. 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. This standard specifies various critical steps in report development that require review and emphasizes that the review process should be conducted according to a schedule prepared well before 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 with the need to make data available, all study procedures must be 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, while the training standard outlines IEA's obligations regarding support for researchers in analyzing study data.

Standard for Conducting Psychometric and Statistical Analyses

Purpose:

To ensure that the study data are modeled, analyzed, and interpreted with precision and accuracy following the analysis plan.

Standard:

Psychometric and statistical analyses should be conducted using the procedures and methods specified in the analysis plan while adhering to established best practices. The implementation process should include careful attention to detail and focus on transparency and reproducibility in the analysis procedures.

Guidelines:

- The international study center should develop detailed specifications for applying the psychometric and statistical methods in the analysis plan and have them reviewed by IEA's Technical Executive Group.
- Implementing psychometric analyses involves thoroughly reviewing the data collected on the assessment instruments. Project teams should work closely with psychometric experts to examine empirical evidence supporting the reliability and validity of the instruments, ensuring that they consistently measure the intended constructs within and across countries. This collaborative effort strengthens the credibility of the psychometric analysis outcomes.
- Statistical analysis involves systematically applying the chosen statistical procedures to the international database. Project teams should carefully consider the potential impact of outliers, missing data, and other pertinent factors that may influence the results. Attention to detail is paramount for producing robust and credible outcomes.
- Throughout the analysis phase, project teams should consult with technical experts as necessary, fostering a collaborative environment that enhances the robustness of the analysis. This thorough review approach ensures that the analysis plan is faithfully executed and scrutinized by experts who bring valuable insights and expertise. This ensures that the entire analytical framework is open to scrutiny and review, contributing to the overall dependability of the findings.

Standard for Developing Sampling Weights

Purpose:

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

Standard:

A sampling weight should be assigned to each sampled unit. This should reflect the probability of selection of the unit. Then, for responding units weight adjustments should be applied to adjust for differential nonresponse, and possibly to reduce sampling variance.

Guidelines:

- Every unit selected in the sample should have a sampling weight, often referred to as a base weight, that reflects its probability of selection in the sample. Responding units in the sample are likely to have one or more additional components: (i) a correction factor to reduce nonresponse bias by reflecting the participation rate among eligible sampled units in the stratum to which the sampled unit belongs; (ii) a trimming factor that reduces the impact of large weights; and (iii) an adjustment that aligns the distribution of the sample with that of known population characteristics, often referred to as a calibration adjustment. One specific such adjustment that is commonly used is known as poststratification. Note that using external information about the population can lead to biases if such information is not accurate regarding the study population (e.g., is out-of-date, uses different definitions, does not cover the same population).
- In some circumstances, it may be desirable to implement nonresponse adjustments to modify the weight of each responding unit such that the weighted distribution of the responding sample on key characteristics matches that of the full weighted sample of eligible units, or, in some circumstances, the full eligible population. Post-stratification might be considered if feasible.
- 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 these weighting components, plus any other adjustment components, from all sampling stages should be produced for each element in the sample. The arising “final” weight variable should be used for parameter estimation.
- 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.
- 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.
- The distribution of the weights among the sampled elements should be checked for 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 as well as non-sampling error should be presented for all statistics included in IEA reports. Data quality measures such as response rates, population coverage and exclusions, also should be reported.

Guidelines:

- Sampling errors 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.
- Sampling error computation should reflect the complexities of the sampling design. Software incorporating methods accounting for the complex sampling design, such as jackknife or balanced repeated replication, can be readily developed for many purposes and has been extensively used in various contemporary large-scale assessments in education with great success. Where appropriate and available, the effects of other error sources (e.g., measurement or linking error) should be incorporated with the effects of sampling error.
- The data files and accompanying documentation should provide sufficient information to permit the calculation of sampling errors for a wide range of statistics.
- Sampling participation rates should be calculated for each country at each stage of sampling and combined across all stages.
- 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 Measurement Scales

Purpose:

To ensure that all measurement scales reported by IEA studies provide defensible evidence that the measures they produce support reasonable interpretations of their intended purpose.

Standard:

The validity of all scales reported in IEA studies should be established and documented. Content validity should be established through the implementation and documentation of rigorous instrument/item development. Evidence of construct validity should be established by evaluating the degree to which the association between selected variables is consistent with expectations based on understandings of the constructs and existing literature. Indices of measurement equivalence should be used to evaluate the consistency of the internal structure of scales across countries and across subgroups of interest. These processes and outcomes should be documented in the technical report, but also in other study reports where appropriate. International reports should include evidence of the reliability of all reported scales.

Guidelines:

- Content validity is an essential precondition to establishing that valid interpretations can be made from measures of achievement and of affective and behavioral constructs. Content validity should be addressed by ensuring that:
 - a) the assessment framework has been certified by leading subject matter experts as a satisfactory representation of the relevant content; and
 - b) the items (and instruments) are developed using a rigorous process with input from experts in the content areas of the field and experts across countries providing feedback on the suitability of content across countries.
- Evidence that scales are valid measures of the intended construct should be established using statistical techniques that show the relationships between the scales and other related measures. With respect to achievement scales, this may be done, for example, by establishing whether the association between the scaled measure of achievement and other variables is consistent with expectations based on research literature and expert understanding of the domains. When considering affective and behavioral domains, similar comparisons between the observed data and expectations may be established through techniques such as factor analysis and structural equation modelling.
- Analyses of achievement and context scales should be conducted with measurement models that make testable assumptions. These models enable verification of the assumptions made, and whenever measurement scales are constructed using psychometric models such as item response theory that are based on strong assumptions, it is crucial to verify that the data adhere to the model assumptions and exhibit appropriate model-data fit. Deviations from model assumptions and poor fit can significantly impact the accuracy of the resulting measures of internal scale structure including reliability, and measurement invariance.
- The equivalence of the measurement properties of scales (measurement invariance) across countries and groups of interest should be investigated as much as possible before reporting results. This may be done using the multigroup confirmatory factor analysis model, or the multiple-causes multiple-indicators model or by calculating standardized effect size measures that evaluate the magnitude of measurement invariance. See von Davier, Gonzalez, and Schultz (2020) for a discussion of measurement invariance of achievement measures in the context of international large-scale assessment.
- An internal consistency index, such as Cronbach's Alpha (or the equivalent measure relevant to the scaling model being applied) should be used to provide evidence of the reliability of scales. Other reliability indices, based, for example, on factor analysis models, should also be calculated and reported whenever possible. Scales exhibiting low reliability should be identified in reports and interpreted with caution.

Standard for Presenting Study Findings

Purpose:

To facilitate widespread use of the study results and ensure the valid interpretation of findings aligned with the study's intended purposes (see also Standard for Reviewing the Primary Reports of Study Findings).

Standard:

IEA studies should present findings clearly, accurately, and in a timely fashion, with indicators of statistical uncertainty as appropriate.

Guidelines:

- The international study center should develop a publication and reporting plan that addresses the types, delivery modes, and numbers of publications to be produced, the process for developing and disseminating the reports, and the production schedule. The analysis and reporting activities should be scheduled to ensure timely availability of the publications.
- All efforts should be made to maximize the use of information and communication technology to promote awareness of and interest in the study findings. This should include study websites that provide access to a range of attractive and accessible content including dynamic multimedia displays of study findings.
- Study findings can be presented as, for example, executive summaries, infographics, general audience reports, innovative assessment tasks, news release videos, dynamic configurable data displays, articles and brochures, and technical reports.
- All results should be accurately reported and made available in tables, graphs, or plots, in a format that is best suited to the data. Data and variable labels should be clear and concise. Any statistics, including derived summary statistics and model based estimates should be reported with standard errors or confidence intervals. The methods for estimating standard errors should be described, and it should be explained what assumptions were made.
- Dynamic configurable data displays can be a useful way to encourage users to engage with the data.
- Best practices and international standards should be used in representing estimated statistics, the uncertainty of the estimates, and in guiding inferences either through statistical testing, confidence or credible intervals, or other statistical tools.
- Graphic displays should be clear, simple to read, and appropriate for the data displayed. Axes and axis labels for charts and figures should be configured to suit the data being reported. Avoid using histograms with zero points for achievement scales that have no meaningful zero (e.g., scales on the 500,100 metric).

Standard for Reviewing the Primary Reports of Study Findings

Purpose:

To ensure that primary study reports accurately and effectively present the most important findings in alignment with the study's objectives.

Standard:

The publication plan and primary study reports should undergo comprehensive reviews by relevant constituencies and advisory committees at key stages, including the report outline, proposed tables and graphs, text drafts, and the final report. The review process should be as inclusive as possible to accommodate a wide range of perspectives. A review schedule, established well before the publication deadline, should guide the review process, with all involved parties given adequate notification.

Guidelines:

- The publication plan should be reviewed by all interested parties, including content matter advisory committees and advisory groups, the IEA Technical Executive Group, and Publications and Editorial Committee, as well as by national research coordinators and their advisors.
- Report outlines and proposed table shells should be reviewed by participants early in the report-writing process, so that changes may be incorporated as necessary.
- National center representatives should be given an opportunity to review their data in summary form to ensure accurate data for each country. The international study center should review all achievement and questionnaire data for accuracy and consistency across countries before cross-country analyses and incorporation in report tables and graphs.
- The international study center should continuously check for accuracy as tables and graphs are updated during the report review process.
- National research coordinators should have an opportunity to review drafts of the reports, including both data tables and text, before these are finalized.
- The Publications and Editorial Committee should review the final drafts.

Standard for Releasing Data

Purpose:

To ensure that international study data are released under controlled and agreed upon procedures.

Standard:

At the beginning of the study, a policy should be established covering data release and confidentiality agreements. The policy should cover preliminary versions of the data as well as the final international database.

Guidelines:

- The international study center should prepare a release policy defining which parties (participating countries, secondary analysts, etc.) will get which data, when, and in what format. The policy should include the requirement that national center staff release their country data and maintain data confidentiality. Agreement with this policy should be secured in writing from all participants as part of the contract between IEA and the participating countries.
- National center staff should have an early opportunity to discuss and agree on timelines for the data release and the format for the data release.
- All released data should be documented at the time of release. Log and process data should be provided with documentation of the data saving logic, documentation of the machine scoring rules and annotated screenshot, unless confidentiality prevents the publication of this information.
- At the end of the study, the international database should be made available as Public Use Files and as Restricted Use Files, in line with legal requirements for data confidentiality. These files should be released with a user guide and sufficiently detailed documentation to permit a user without previous familiarity with the study to analyze the data. This should be accompanied by an updated version of the IDB Analyzer, IEA's statistical analysis software for analyzing complex data.
- The international database should be distributed in a format accessible to the widest possible audience, avoiding reliance on data formats that require proprietary software for access.

Standard for Preparing Technical Reports and Documentation

Purpose:

To ensure that all major technical aspects of the study are clearly and completely documented.

Standard:

The study design, instrument development, data collection, and analysis and reporting procedures should be described in sufficient detail to permit them to be evaluated and replicated.

Guidelines:

- Technical reports should describe in detail the design and implementation of the study, including sampling design, instrument development and data collection, quality assurance activities, platforms used for implementation, data cleaning and processing, and psychometric analysis and outcomes.
- The technical documentation should be produced promptly, ideally simultaneously with the publication of the first international reports. As IEA studies extend over several years, it can be helpful to produce technical report chapters at critical points of the study and publish them online as the study progresses.
- The development of achievement and questionnaire items should be described in detail. It should be clear how measurement of trends was maintained during instrument development.
- Sampling implementation in each country should be described, including accurate information on population coverage, participation rates, and sampling precision. Countries not reaching study standards for coverage or participation should be clearly annotated.
- Technical reports should clearly document the quality assurance procedures used at 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.
- Procedures for scoring or quantifying responses to achievement or questionnaire items should be described for both human and machine scoring activities, together with the steps taken to ensure accuracy and consistency across scorers.
- Full documentation should be provided about the data collection technology, processing, and storage of data.
- Advanced or innovative techniques for scaling or analyzing study data should be documented, and evidence of their efficacy provided if such information is not already available in the public domain.
- All error terms and the methods used to estimate the errors should be clearly described.
- At the end of the study, a fully documented database is made available to the public. The documentation of analytic procedures should be sufficiently detailed to permit someone with access to the data to replicate the results.

Standard for Ensuring Data Integrity

Purpose:

To ensure the integrity of data to support all subsequent analyses for accuracy and comparability within a country, across countries, and over multiple cycles.

Standard:

The data provided by national centers should include a verifiable chain of custody, with time stamps, and adhere to the prescribed responses and data structure in the code book. At each major data transfer, specified standardized analyses should be conducted to identify any unexpected behavior suggested by the data at the item level as well as across sets of items. Where feasible, comparisons should be made to data associated with response accuracies and timing collected in previous cycles or between the field trial and main study. The source of any unexpected data should be investigated.

Guidelines:

For the International Study Center:

- The international study center should provide participating countries with clear and verifiable procedures for all steps of the data process chain and to ensure the integrity and confidentiality of the data. This applies to all processes in both paper-based assessment and technology-based assessment.
- All IT systems set up and maintained by the international study center should therefore require secure and reliable authentication, store the data encrypted if possible and protected against unauthorized access from outside, carry out secure data transfer via encrypted connections if possible, and log all changes to the data.
- The raw responses given by the students should be kept as a reference for later comparisons or reviews. For this purpose, test booklets should be kept in the country at least until the report is submitted. In the case of technology-based assessment, the raw responses should be archived immediately after receipt of the data (i.e., the data upload).
- The international study center should employ “best practice” procedures for detecting irregularities in assessment data.

Clear differences between machine and human scored data can also indicate issues. Any such data should be interpreted in comparison to similar data collected in other countries in the study.

For National Study Centers

- National study centers must follow all the guidelines of the international study center.
- In particular, national centers should ensure access control for all material used in the assessment and protect it against unauthorized access.
- In paper-based assessments, scorers and keypunchers should immediately inform the national research coordinator if they suspect that student answers have been manipulated.

National research coordinators should immediately inform the international study center if they detect any anomalies in their data or data collection or data processing procedures that they feel may require further investigation.

Standard on Training Primary and Secondary Users of IEA Data

Purpose:

To ensure that users of IEA data are provided sufficient information to support them to deal with the complexities of IEA study data.

Standard:

International study centers should provide technical documentation and training to primary and secondary data users, where statistical analysis techniques appropriate to the study design are described, and sample analysis procedures applying these techniques are presented.

Guidelines:

- IEA assessment data can have a very complex structure and typically require training to analyze the data effectively. Users of the data must be made aware of these complexities and of the analytic procedures available to deal effectively with them.
- Before the release of the international database, the international study center should conduct at least one database training session for national center staff involved in creating national reports.
- In preparation for training, the international study center should prepare a user guide for the international database describing the methods and procedures for analyzing the data as well as a technical report documenting the design, development, implementation, analysis, and reporting of the study.
- This training should explain the structure of the international database, how to merge and manage datasets, and how to use the IEA IDB Analyzer to conduct univariate and multivariate analyses with study data.
- Shortly after the release of the data, the international study center should conduct database training for secondary users of the database.
- The international study center should collaborate with IEA in providing and updating statistical analysis software such as the IDB Analyzer that enables analysts to deal effectively with the complex data structures of IEA studies.

References

- Addey, C., Sellar, S., Steiner-Khamisi, G., Lingard, B. & Verger, A. (2017) The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*, 47(3), 434–452, <https://doi.org/10.1080/03057925.2017.1301399>
- Dolan, R.P., Burling, K.S., Rose, D., Beck, R., Murray, E., Strangman, N., Jude, J., Harms, M., Way, W.D., Hanna, E. and Nichols, A., (2010). *Universal design for computer-based testing (UD-CBT) guidelines*. Pearson. <https://doi.org/10.13140/RG.2.2.31923.35361>
- Hernández-Torrano, D., & Courtney, M.G.R. (2021). Modern international large-scale assessment in education: an integrative review and mapping of the literature. *Large-scale Assessments in Education*, 9(17). <https://doi.org/10.1186/s40536-021-00109-1>
- Martin, M. O., Rust, K., & Adams, R.J. (1999). *Technical Standards for IEA Studies*. International Association for the Evaluation of Educational Achievement (IEA).
- Mullis, I.V.S., & Martin, M.O. (2022). IEA's TIMSS and PIRLS: Measuring Long-Term Trends in Student Achievement. In T. Nilsen, A. Stancel-Piątak, J-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education*. Springer International Handbooks of Education. Springer, Cham. https://doi.org/10.1007/978-3-030-88178-8_15
- World Bank. (2023). *Using Principles of Universal Design for Assessment (UDA) Design Accessible Learning Assessments - Toolkit (English)*. World Bank Group. <http://documents.worldbank.org/curated/en/099100223135027487/P1713680054b4908a0be920cbeb9acd2b2e>

Further Reading

Clarke, M., & Luna-Bazaldua, D. (2021). *Primer on large-scale assessments of educational achievement*. World Bank Publications.

Nilsen, T., Stancel-Piątak, A., & Gustafsson, J-E. (Eds.). (2022). *International handbook of comparative large-scale studies in education: Perspectives, methods and findings*. Springer Nature.
<https://doi.org/10.1007/978-3-030-88178-8>

Rocher, T., & Hastedt, D. (2020, September). *International large-scale assessments in education: A brief guide*. IEA Compass: Briefs in Education No. 10. IEA. <https://www.iea.nl/publications/series-journals/iea-compass-briefs-education-series/september-2020-international-large>

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press.

Wagemaker, H. (Ed.). (2020). *Reliability and Validity of International Large-scale Assessment: Understanding IEA's Comparative Studies of Student Achievement*. Springer, Cham. <https://doi.org/10.1007/978-3-030-53081-5>

This publication presents IEA's *Technical Standards for International Large-Scale Assessment*. The initial standards, published in 1999, aimed to consolidate the best practices and methodological rigor in IEA's approach to educational assessment, addressing the unique needs of international studies. The standards presented in this publication reflect IEA's ongoing commitment to maintaining the highest levels of quality in IEA's studies and contributions to the field of international large-scale assessment (ILSA).

Since 1999, many changes have taken place in the field of ILSA, including in IEA studies. The 2025 standards were built on the foundations of the 1999 standards and have been revised and expanded to accommodate these changes. The four flagship IEA studies—TIMSS, PIRLS, ICCS, and ICILS—are computer-based and cyclical, which offers new opportunities but also new challenges. In addition, since 1999, processes, techniques, and methodological approaches have evolved and contributed to all aspects of study design, operationalization, implementation, analysis, and reporting. The evolution of ILSA, including IEA studies, has necessitated revision of existing standards and, in some cases, the development of new standards to guide their design and implementation. The 2025 standards have been published as a resource to support the many stakeholders in international research, from those directly involved in study implementation to those seeking to use ILSA data and evaluate and refine ILSA as part of a process of continuous improvement.