# CHAPTER 7

# Creating the PIRLS 2021 International Database

Mark Cockle

## Overview

The PIRLS 2021 International Database is available to researchers, analysts, and any individuals interested in the data collected and analyzed as part of PIRLS 2021. The database includes student achievement data, as well as student, home, teacher, school, and national context data, for 57 countries and 8 benchmarking participants. Preparing the PIRLS 2021 International Database and ensuring its integrity was a major undertaking, requiring extensive collaboration among IEA Hamburg, the TIMSS & PIRLS International Study Center, Statistics Canada, and the national centers of participating countries. In creating a large database that contains internationally comparable data collected across a diverse group of countries, languages, cultures, and educational contexts, it is important that standardized procedures are followed and that both the process and content are well-documented for secondary users of the data. After countries prepare and submit their national data and documentation, data processing by IEA Hamburg includes a uniform cleaning procedure for structure, identification variables, linkages, and context data. Before the international database is made available to the general public, three interim versions of the national database files are provided to each country for checking and confirming their data are ready for analysis by the TIMSS & PIRLS International Study Center.

For each PIRLS assessment, participating countries follow standardized survey operations procedures to prepare instruments, administer the achievement test, collect context questionnaire data, and prepare data files and documentation to submit to IEA Hamburg (see Chapter 4). Once the team at IEA Hamburg received the PIRLS 2021 data and documentation, they followed standardized international data management and verification procedures to check for errors and inconsistencies and create standardized data and documentation outputs for each country. These procedures addressed unique aspects for each instrument (achievement test and context questionnaires) and accounted for adaptations to the context questionnaires made for the particular contexts in the participating countries (see Chapter 5).


https://doi.org/10.6017/lse.tpisc.tr2103.kb3876

The international data management process implemented by IEA Hamburg includes verification of the following:

- All information in the database conforms to the internationally defined data structure.
- The content of all codebooks and documentation appropriately reflects national adaptations to questionnaires.
- All variables used for international comparisons are in fact comparable across countries (after harmonization, where necessary).
- All institutions involved in the process apply quality control measures throughout in order to assure the highest quality and accuracy of the data.

To maintain consistency across PIRLS assessment cycles in the interest of measuring trends, IEA Hamburg applies the same general procedures in each cycle while updating some processes to improve efficiency and to accommodate changes to the assessment design, instruments, or other circumstances. In particular, the complexities introduced by the transition to digital assessment necessitated adaptations to procedures in PIRLS 2021. Although some efficiencies were afforded by the transition to digital assessment, such as reduced manual data entry by country participants, additional work was required to develop and validate data handling procedures and to ensure consistency of the data captured across modes of administration. About half of the PIRLS 2021 countries transitioned to digital assessment, and half administered on paper as in previous cycles. This required a process to accommodate data from both digital and paper administrations in a single database with a common international data structure. PIRLS 2021 also required a separate "bridge" database for the paper-based data collected in countries that transitioned to the digital assessment. The bridge data were used to link the paper and digital assessment results onto a common scale (see Chapter 10).

Moreover, in the face of disruptions to teaching and learning worldwide due to the COVID-19 pandemic, many countries had to delay data collection activities, resulting in a greater amount of time having passed between activities. Records for participating students and their parents, teachers, and school principals were carefully tracked, particularly in cases when data collection, scoring, or data entry was postponed due to pandemic-related disruptions to operations at schools or national centers.

For PIRLS 2021, IEA Hamburg was responsible for checking the data files submitted by each country, applying standardized data cleaning rules to verify the accuracy and consistency of the data, and documenting any deviations from the international file structure. For data from the digital PIRLS administration, this included processing and cleaning the data collected by the PIRLS Player software that delivers the digital assessment to students, importing student achievement response data for human scoring into IEA's CodingExpert system, and implementing machine scoring rules for achievement items according to specifications from the TIMSS & PIRLS International Study

Center. Efforts to ensure accuracy and consistency of the digital data began during the national PIRLS Player checking process, which involved extensive, semi-automated data saving checks for all achievement items and student questionnaire items in PIRLS 2021 (see Chapter 5).

National Research Coordinators (NRCs) from each participating country collaborated with IEA Hamburg to resolve any queries that emerged during the data cleaning process and checked any interim versions of their database(s) produced by IEA Hamburg during this process. The TIMSS & PIRLS International Study Center provided NRCs with univariate data almanacs containing summary item statistics on each variable so that the national centers could evaluate their data from an international perspective (see Chapter 9).

The TIMSS & PIRLS International Study Center also conducted all operational psychometric analyses of the achievement and context questionnaire data and produced reading achievement scores (plausible values—see Chapter 11), context questionnaire scores (Chapter 15), and other derived variables based on the context data. Using the Within-School Sampling Software (WinW3S)[1] database and achievement data provided by IEA Hamburg, Statistics Canada in collaboration with IEA Hamburg calculated the sampling weights, population coverage, and school and student participation rates (Chapter 8).

The *PIRLS 2021 User Guide for the International Database* describes all data files and their variable contents, along with documentation about the achievement items and context questionnaire items. A supplement to the user guide provides the National Adaptations documentation for the PIRLS 2021 Context Questionnaires.

## Preparing and Submitting National Data and Documentation

Data collected as part of PIRLS 2021 required work by participating countries before being submitted to IEA Hamburg for processing and cleaning. After processing and cleaning, data went to the TIMSS & PIRLS International Study Center for verification and analysis and to Statistics Canada for calculating sampling weights and outcomes. This included data collected from: 1) the PIRLS Player, which delivered the digital assessment and student context questionnaire; 2) paper instruments, including paper achievement booklets and context questionnaires; and 3) IEA's Online Survey System, which countries could use to administer home, teacher, and school questionnaires.

### Data from DigitalPIRLS Administration

The digitalPIRLS assessment was designed to run online, through web delivery, or locally, on PC devices using USB delivery. Whereas web delivery ensured that data were immediately available on servers for further processing, USB delivery needed test administrators to upload the student

1 WinW3S is a software developed by IEA Hamburg that allows users to perform all necessary within-school sampling activities according to the PIRLS standards, and also provides some data validation in and across these levels. The software stores participation information at school, teacher, class, and student levels in a relational database, necessary for calculating participation rates and sampling weights.

response databases after the testing session. National centers could use a data monitoring tool that listed all student records present on the upload server and allowed for downloading the list to be used for checking and updating the data availability status in WinW3S.

### Pre-Processing and Scoring Digital Data

Some pre-processing steps were required to prepare digital data in a suitable format for scoring and further processing. Data from web-based records or USB uploads from the PIRLS Players were received daily at IEA Hamburg via SFTP transfer. These data, in the form of JSON (JavaScript Object Notation) files (one file per digital "booklet" and per country/language combination), were then converted into a SQL structure for further processing. This new structure contained a separate database for each country and language, including all data from the original file, identification variables related to the import of data, and additional fields for scoring purposes.

Student responses to constructed response items requiring scoring by humans were transferred to the online IEA CodingExpert system to be allocated to national teams of scorers. At the start of this process, it was essential that the scoring system did not contain any duplicate records including identical responses from the same student. In addition to measures that prevented a database from being uploaded a second time from the client side, checks were made to the student response database to ensure any possible duplicates were identified and reconciled before import. Once IEA Hamburg applied data processing procedures to merge any incomplete student records and resolved any remaining issues with duplicates, the scoring supervisors distributed responses to the scorers on their scoring teams. The IEA CodingExpert software was used by NRCs and their scoring staff to score the digital constructed response items. When scoring was completed, the student response data were transferred to tables prepared for import into the data processing system employed at IEA Hamburg for all large-scale international assessments.

## Data Entry and Verification of Paper Instruments

Each national center was responsible for entering the responses collected in paper achievement booklets and paper-based context questionnaires into data files using the IEA Data Management Expert (DME) software. DME is a software system developed by IEA Hamburg that facilitates data entry and includes validation checks to identify inconsistencies. National centers were instructed to enter data for any questionnaire that contained at least one valid response and to discard unused or empty instruments. This applied to countries that administered the paper assessment as well as the digital assessment, as these countries administered at least some questionnaires on paper and also administered paper bridge booklets of achievement items to a smaller sample of schools and students.

National centers entered responses from the paper instruments into data files using a predefined international codebook. The codebook defines the structure of the data to be entered

and contains information about the names, lengths, labels, and missing codes of variables; valid response ranges for continuous measures or counts; and valid values for nominal or ordinal questions.

As described in Chapter 5, countries participating in PIRLS are expected to make national adaptations to certain questions in the international questionnaires (e.g., the questions about parents' education must be adapted to the national context). Countries making such adaptations were required to update the codebook structure to reflect the adaptations made to the national questionnaire versions before beginning the data entry process.

To ensure consistency across participating countries, the basic rule for data entry into DME required national staff to enter data "as is" without any interpretation, correction, truncation, imputation, or cleaning.

The guiding principles for data entry included the following:

- Responses to closed response items were coded as "1" if the first option was used, "2" if the second option was marked, and so on.
- Responses to open response questions, for example number of students in the sampled class, were entered "as is" even if the value was outside the originally expected range.
- Responses to filter questions and filter-dependent questions were entered exactly as filled in by the respondent, even if the information provided was logically inconsistent.
- Non-response, ambiguous responses, responses given outside of the expected format, or conflicting responses (e.g., selection of two options in a multiple-choice question) were coded as "omitted."

As each respondent ID number was entered, it was checked by the DME software for alignment with a five-digit checksum generated by WinW3S. A mistype in either the ID or the checksum resulted in an error message prompting the person entering the data to check the entry. The data verification module of the DME also checked for a range of other issues such as inconsistencies in identification codes and out-of-range or otherwise invalid codes. When such issues were flagged by the software, the data entry staff were prompted to resolve the inconsistency before resuming data entry.

## Double-Data Entry

To check data entry reliability in participating countries, national centers were required to have a 5 percent sample of each survey instrument (achievement booklet or questionnaire) entered a second time by a different data entry person, operating independently from the first. IEA Hamburg recommended that countries begin the double-data entry process as early as possible during the data capture period in order to identify possible systematic misunderstandings or mishandlings of data entry rules and to initiate appropriate remedial actions—for example, retraining national center staff.

Although it was desirable that every discrepancy be resolved before submission of the complete dataset, the acceptable level of disagreement between the originally entered and double-entered data was established at 1 percent or less for questionnaire data and at 0.1 percent or less for achievement data. Values above these levels required resolution of the discrepancy and re-entry of data.

The level of disagreement between the originally entered and double-entered data was evaluated by IEA Hamburg, and it was found that in general the margin of error observed for processed data was well below the required thresholds.

## Data from Online Questionnaire Administration

National centers had the option of administering the school, teacher, and home questionnaires online through IEA's Online SurveySystem instead of or in addition to using paper-based questionnaires. In addition, NRCs from participating countries completed the PIRLS 2021 Curriculum Questionnaire through this system.

To ensure confidentiality, national centers provided every respondent a letter containing individual login information along with information on how to access the online questionnaire. This login information corresponded to the ID and checksum provided by WinW3S. This embedded the identity validation step into the questionnaire login process, rather than requiring that validation occur at the national center, as was the case with paper-based data entry.

Online administration of questionnaires had a number of advantages. Because responses were collected in digital format and stored directly on the IEA Hamburg server, there was no need for data entry, reducing the workload for national centers. Also, because the online system did not allow for inconsistent response patterns, the data collected had fewer inconsistencies than data collected through the paper-based questionnaires. For example, if the directions ask the respondent to "Check one circle for each line," the system did not allow the respondent to check more than one response category on each line.

The PIRLS 2021 online questionnaires also included skip logic, which minimized response burden and improved data consistency. PIRLS questionnaires have a number of questions that filter out respondents—meaning the subsequent questions are not applicable given the response to the filter question. For example, Question 7A of the school questionnaire reads "Does your school have a school library? If No, go to #8." If a respondent chooses "No," the online survey skips directly to Question 8, omitting Question 7B. Not only did the skip logic save respondents' time, but it also resulted in fewer inconsistencies in the data received by IEA Hamburg and instead produced planned missingness of the skipped responses which were coded in the final database as "not applicable."

### Data Verification at the National Centers

Before sending the data to IEA Hamburg for further processing, national centers carried out mandatory validation and verification steps on all entered data and undertook corrections as necessary.

While the questionnaire data were being entered, the data manager or other staff at each national center used the information from the tracking forms to verify the completeness of the materials. Student participation information (e.g., whether a student participated in the assessment or was absent) was entered via WinW3S.

The validation process was supported by an option in WinW3S to generate an inconsistency report. This report listed all of the types of discrepancies between variables recorded during the within-school sampling and test administration processes and made it possible to cross check these data against data entered in the DME, the database for online respondents, and the uploaded student data on the central international server.

Data managers were requested to resolve such issues before final data submission to IEA Hamburg. If inconsistencies remained or the national center could not solve them, IEA Hamburg asked the center to provide documentation on these problems.

Upon submitting the validated data to IEA Hamburg, NRCs also provided extensive documentation, including hard copies or electronic scans of all original Student and Teacher Tracking Forms, Student Listing Forms, and when applicable, a report on procedural activities collected as part of the online Survey Activities Questionnaire (see Chapter 4).

## Data Processing and Data Cleaning

To ensure the integrity of the international database, IEA Hamburg followed uniform data cleaning procedures as part of processing national data, involving regular consultation with NRCs. The main objectives of the data cleaning process were to ensure that the data adhered to international formats, that school, teacher, and student information could be linked across different survey files, and that the data reflected the information collected within each country in an accurate and consistent manner.

After each country had submitted its data, codebooks, and documentation, IEA Hamburg, in collaboration with the NRCs, conducted a four-step cleaning procedure upon the submitted data and documentation:

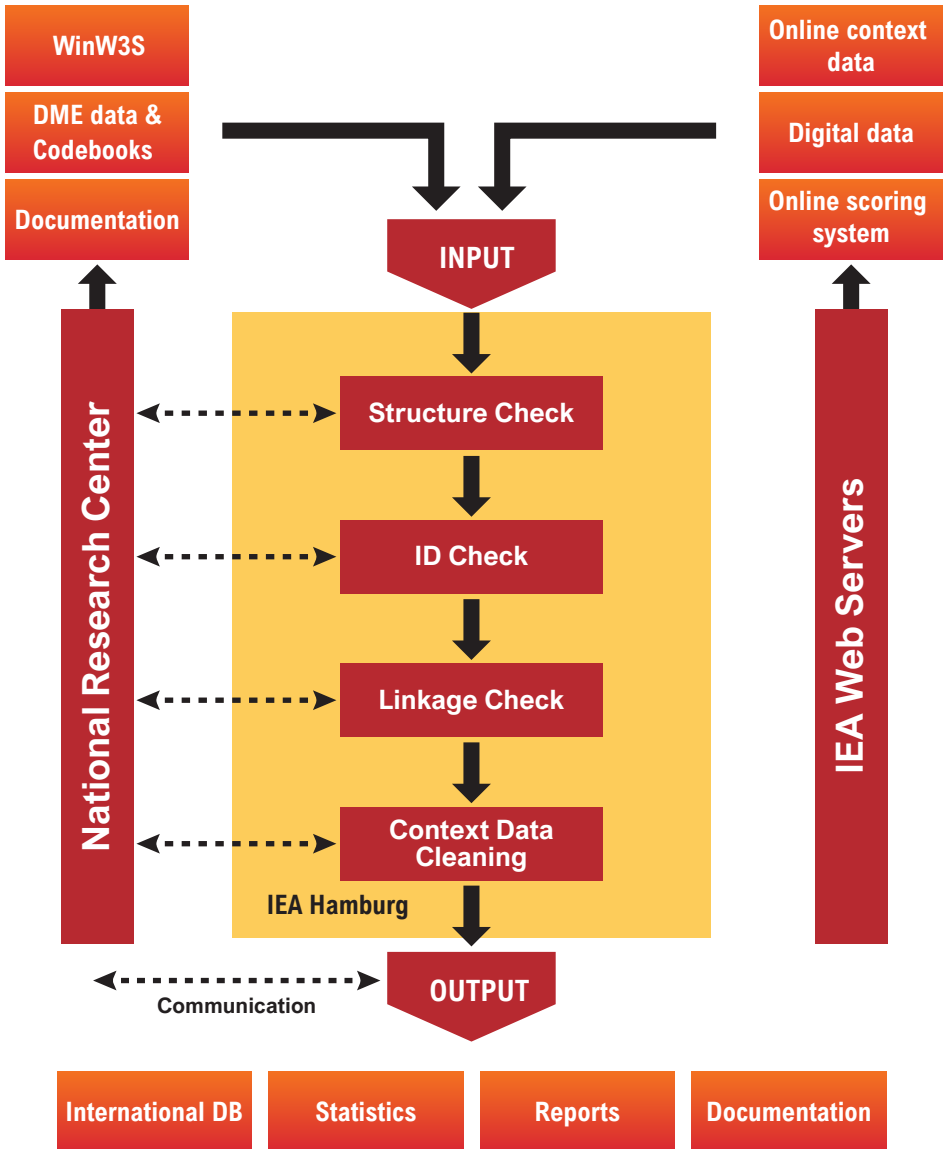
- Checking structure and documentation
- Checking identification (ID) variables
- Checking linkages
- Context data cleaning

Data processing included numerous iterations of the four-step cleaning procedure and were completed on each national data set in close collaboration with national centers. This iterative multi-step cleaning ensured that all data were properly cleaned and that any new errors that could have been introduced during the data cleaning were rectified. Any inconsistencies detected during the cleaning process were resolved in collaboration with national centers, and all corrections made during the cleaning process were documented in a cleaning report, produced for each country. The cleaning process was repeated as many times as necessary until all data were made consistent and comparable.

After the final cleaning iteration, each country's data were sent to Statistics Canada for the calculation of sampling weights. Then the data, including sampling weights, were sent to the TIMSS & PIRLS International Study Center so that the psychometric analyses (as described in Chapter 9 and Chapter 11) could be conducted. The NRCs were provided with interim data products to review at different points in the process.

As illustrated in Exhibit 7.2, the program-based data cleaning consisted of a set of activities explained in the following subsections. IEA Hamburg carried out these activities in close communication with the national centers, as well as with the TIMSS & PIRLS International Study Center.

**Exhibit 7.2:** Overview of Data Processing at IEA Hamburg




## Checking Structure and Documentation

For each country, data cleaning began with a review of data file structures and its data documentation, including a review of National Adaptation Forms, Student Tracking Forms, Teacher Tracking Forms, Student Listing Forms, and the Survey Activities Questionnaire.

After the review, IEA Hamburg first merged the tracking information and sampling information captured in the WinW3S database with the student-level database containing the corresponding student data from either paper or digital achievement assessments. During this step, IEA Hamburg

staff also merged the data from the school, home, and teacher questionnaires for both the online and paper modes of administration. At this stage, data from the different sources was transformed and imported into one SQL database so that this information would be available during all further data processing stages.

The first checks identified differences between the international and the national file structures. Some countries made adaptations (such as adding national variables or omitting or modifying international variables) to their questionnaires. The extent and nature of these changes differed across countries: some countries administered the questionnaires without any modifications (apart from translations and necessary adaptations relating to cultural or language-specific terms), whereas other countries inserted response categories within existing international variables or added national variables.

To keep track of adaptations, staff at the TIMSS & PIRLS International Study Center asked the national centers to complete National Adaptation Forms. In their adaptations, countries sometimes modified the structure and values of the international codebooks. In such cases, IEA Hamburg had to recode variables in the national data files to ensure that the resulting data remained comparable across countries. The national adaptation process is described in Chapter 5 and details about country-specific adaptations to the international instruments can be found in Supplement 2 of the *PIRLS 2021 User Guide for the International Database*.

IEA Hamburg then discarded variables created purely for verification purposes during data entry and made provision for adding new variables necessary for analysis and reporting, including reporting variables, derived variables, sampling weights, and scale scores.

Once IEA staff had ensured that each data file matched the international format, they applied a series of standard data cleaning rules for further processing. Processing during this step employed software checks developed by IEA Hamburg to identify and correct inconsistencies in the data. Each potential problem flagged at this stage was identified by a unique problem number and then described and recorded in a database. The action taken by the cleaning program or IEA Hamburg staff with respect to each problem was also recorded.

IEA Hamburg referred problems that could not be rectified automatically to the responsible NRC so that national center staff could check the original data collection instruments and tracking forms to trace the source of these errors. Wherever possible, staff at IEA Hamburg suggested a remedy and asked the national centers to either accept it or propose an alternative. If a national center could not solve the issue through verification of the instruments or forms, IEA Hamburg applied a general cleaning rule to the files to rectify the error. When all automatic updates had been applied, IEA Hamburg staff used SQL recoding scripts to directly apply any remaining corrections to the data files.

## Checking Identification Variables

Each record in a data file needs to have a unique identification number. The existence of records with duplicate ID numbers in a file implies an error of some kind. Some countries administered the school, teacher, and home questionnaire online in addition to the paper mode. Therefore, by mistake a respondent could have completed both the paper and the online versions of the questionnaire. Similarly, it was possible for a digital assessment login to be used (and responses uploaded) twice. If two records in a PIRLS 2021 database shared the same ID number and contained exactly the same data, IEA Hamburg deleted one of the records and kept the other one in the database. In the rare cases where records contained different data and IEA staff found it impossible to identify which record contained the more reliable or complete version of the data, national centers were asked which record to keep.

Although the ID cleaning covered all data from all instruments, it focused mainly on the student data files. In addition to checking the unique student ID number, it was crucial to check variables pertaining to student participation and exclusion status. Confirming students' birth dates and dates of testing was also crucial in order to correctly calculate student age at the time of testing. The Student Tracking Forms provided an important tool for resolving anomalies in the database.

## Checking Linkages

As data on students, parents, teachers, and schools appeared in a number of different data files, a process of linkage cleaning was implemented to ensure that the data files would correctly link together. The linking of the data files followed a hierarchical system of identification codes that included school, class, and student components. These codes linked the students with their class and school membership. Further information on linkage codes can be found in Chapter 4.

Linkage cleaning consisted of a number of checks to verify that student entries matched across achievement files, student context questionnaire data files, scoring reliability files, and home background files. In addition, at this stage, checks were conducted to ensure that teacher and student records linked correctly to the appropriate schools. The Student Tracking Forms, Teacher Tracking Forms, and Student Listing Forms were crucial in resolving any anomalies. IEA Hamburg also liaised with NRCs about any problematic cases, providing the national centers with standardized reports listing all inconsistencies identified within the data.

As mentioned previously, IEA Hamburg conducted all cleaning procedures in close cooperation with the national centers. After national center staff had cleaned the identification variables and linkages, IEA Hamburg passed the clean databases with information about student participation and exclusion on to Statistics Canada, which used this information to calculate students' participation rates, exclusion rates, and student sampling weights.

## Context Data Cleaning: Resolving Inconsistencies in Context Questionnaire Data

The amount of inconsistent and implausible responses in questionnaire data files varied considerably across countries. IEA Hamburg determined the treatment of inconsistent responses on a question-by-question basis, using all available documentation to make informed decisions. IEA Hamburg staff also checked all questionnaire data for consistency across the responses given. For example, Question 1 in the school questionnaire asked for the total school enrollment in all grades, and Question 2 asked for the enrollment in the target grade only. Logically, the number given as a response to Question 2 could not exceed the number provided by school principals in Question 1. Similarly, it would not be possible for the number of years a teacher has been teaching altogether (Question 1 in the teacher questionnaire), subtracted from their age (Question 3), to be lower than the minimum possible age of a beginning teacher in their particular country. IEA Hamburg flagged inconsistencies of this kind and then asked the national centers to review these issues. IEA staff recoded as "invalid" those cases that could not be corrected.

Filter questions, which appeared in some questionnaires, directed respondents to a particular set of questions only applicable to a subset of respondents. IEA Hamburg applied the following cleaning rule to these filter questions and its dependent questions: If a respondent answered "No" to Question 7A in the school questionnaire "Does your school have a school library?", but responded to the dependent question 7B about the number of books in the library, IEA Hamburg recoded those responses to 7B as "logically not applicable." Also, following the same example, if the filter question was omitted but at least one valid response was found in the dependent questions, then IEA Hamburg recoded the filter question to "Yes." This of course is only possible for dichotomous filter questions (e.g., with response options such as "Yes/No").

IEA Hamburg also applied what are known as split variable checks to questions where the answer was coded into several variables. For example, Question 5 in the student questionnaire asked students the following: "Do you have any of these things at your home?" Student responses were captured in a set of seven variables, each one coded as "Yes" if the corresponding "Yes" option was filled in and "No" if the "No" option was filled in. Occasionally, students checked some "Yes" boxes in the set but left the rest unchecked. Because, in these cases, it was clear that the unchecked boxes meant "No," these responses were recoded accordingly.

In addition, student reports to items on gender and age in the student questionnaire were checked against the tracking information provided by the School Coordinator or Test Administrator during the within-school sampling and test/questionnaire administration process. When information on gender or birth year and month was missing in the student questionnaire, this information, when available, was copied over from the tracking data to the questionnaire. If discrepancies were found in gender and/or age between student questionnaire responses and existing tracking data, IEA

Hamburg queried the case with the national center, and the national center investigated which source of information was correct. If unresolved, tracking data was used rather than questionnaire data.

## Handling of Missing Data

Two overarching types of entries were possible during the PIRLS 2021 data capture: valid data values and missing data values. During data capture, missing data was assigned a value of "omitted/invalid" or "not administered." IEA Hamburg applied additional missing data codes to facilitate further analyses. This process led to four distinct types of missing data in the international database:

- **Omitted or invalid:** The respondent had a chance to answer the question but did not do so, leaving the corresponding item or question blank. This code was also used if the response was uninterpretable or out-of-range.
- **Not administered:** The item or question was not administered to the respondent, which meant that the respondent could not read and answer the question. The "not administered" missing code was used for those student test items that were not in the set of assessment blocks administered to a student, either deliberately (due to the rotation of assessment blocks), or in rare cases, due to technical failure or incorrect translations. This missing code was also used for those records that were included in the international database but did not contain a single response to one of the assigned questionnaires. For example, this situation applied to home questionnaire data for students who participated in the student test but whose parent/guardian did not answer the home questionnaire. In addition, the not administered code was used for individual questionnaire items that a national center decided not to include in the country-specific version of the questionnaire.
- **Logically not applicable:** The respondent answered a preceding filter question in a way that made the following dependent questions not relevant to him or her.
- **Not reached:** This applied only to the individual items of the student achievement test. It indicated those items that the student did not attempt at the end of the booklet, either because time ran out or because the student stopped responding. "Not reached" codes were derived as follows: First, the last sequential response given by a student in a session was identified. Then, the first response after this last answer was coded as "omitted." Finally, all following responses to these items in the session were coded as "not reached." For example, the response pattern "1 9 4 2 9 9 9 9 9" (where "9" represents "omitted") would be recoded to "1 9 4 2 9 R R R R R" (where "R" represents "not reached")

## Data Cleaning Quality Control

PIRLS 2021 was a large and highly complex study with very high standards for data quality. Maintaining these standards required an extensive set of interrelated data checking and data cleaning procedures. To ensure that all procedures were conducted in the correct sequence, that no special requirements were overlooked, and that the cleaning process was implemented independently of the persons in charge, the data quality control process included the following:

- **Thoroughly testing all data cleaning programs:** Before applying the programs to real datasets, IEA Hamburg applied them to simulation datasets containing all possible problems and inconsistencies.
- **Registering all incoming data and documents in a dedicated database:** IEA Hamburg recorded the date of arrival as well as specific issues requiring attention.
- **Carrying out data cleaning according to strict rules:** Deviations from the cleaning sequence were not possible, and the scope for involuntary changes to the cleaning procedures was minimal.
- **Documenting all systematic data recoding applied to all countries:** IEA Hamburg recorded all changes to data in the comprehensive cleaning documentation provided to national centers.
- **Logging every "manual" correction to a country's data files in a recoding script:** Logging these corrections, which occurred only occasionally, allowed IEA Hamburg staff to undo specific changes or redo the whole manual cleaning process at any later stage of the data cleaning process.
- **Repeating, on completion of data cleaning for a country, all cleaning steps from the beginning:** This step allowed IEA Hamburg to detect any problems that might have been inadvertently introduced during the data cleaning process.
- **Working closely with national centers at various steps of the cleaning process:** IEA Hamburg provided national centers with the processed data files and accompanying documentation so that center staff could thoroughly review and correct any identified inconsistencies.

IEA Hamburg compared national adaptations recorded in the documentation for the national datasets with the structure of the submitted national data files. IEA Hamburg staff then recorded any identified deviations from the international data structure in the national adaptation database and for the supplementary materials provided with the *PIRLS 2021 User Guide for the International Database*. Whenever possible, IEA Hamburg recoded national deviations to ensure consistency with the international data structure.

## Interim Data Products

Before the PIRLS 2021 International Database was finalized, three major interim versions of the data files were sent to each country. The dates when these versions were provided depended on the time countries administered the assessment. Documentation, with a list of the cleaning checks and corrections made in the data, was included with the data files. For the three interim versions, each country only received its own dataset and did not have access to other countries' data. Countries that administered digitalPIRLS received additional files with student raw responses. These raw response files are the trace of what students answered and are in this sense comparable to the completed paper booklets that paperPIRLS countries would have available for checking.

The first version of the interim data files was sent as soon as the data could be considered "clean" in regard to identification codes, linkage issues, and context data inconsistencies. A second version was sent to the countries when all national adaptations and the feedback resulting from the review of the first version were implemented. NRCs were asked to confirm that the data were ready for the operational psychometric analysis used for achievement scaling. A third version of the data files was sent to countries when the weights and achievement plausible values were available and had been merged with the data files. This version, sent to the countries in February 2023, was structurally equivalent to the files to be published in the PIRLS 2021 International Database. It contained only those student records that were used in the analysis by the TIMSS & PIRLS International Study Center and satisfied the sampling standards.

Interim data products were accompanied by detailed data processing and national adaptation documentation, codebooks, and summary statistics. The summary statistics were created by the TIMSS & PIRLS International Study Center and included weighted univariate statistics for all questionnaire variables for each country. For categorical variables, representing the majority of variables, the percentages of respondents choosing each of the response options were displayed. For continuous numeric variables, various descriptive statistics were reported, including the minimum, maximum, mean, median, mode, and percentiles. For both types of variables, the percentages of missing data were reported. Additionally, for the achievement items, the TIMSS & PIRLS International Study Center provided item analysis and reliability statistics listing information such as the number of valid cases, percentages, percentage correct, Rasch item difficulty, and scoring reliability. These statistics were used for a more in-depth review of the data at the international and national levels in terms of plausibility, unexpected response patterns, etc. More information on item almanacs and reviewing item statistics is available in Chapter 9.

## Final Product—the PIRLS 2021 International Database

The extensive data cleaning effort implemented at IEA Hamburg ensured that the PIRLS 2021 International Database contained high quality, internationally comparable data. More specifically, the process ensured that:

- Information coded in each variable conformed to the international scheme
- National adaptations were reflected appropriately in all variables
- All entries in the database could be successfully linked across students, teachers, and schools
- Sampling weights and student achievement variables were available for international comparisons.

The PIRLS 2021 International Database will be made available to the general public on June 22, 2023. Accompanying the data files, the *PIRLS 2021 User Guide for the International Database* describes the structure of the database, the variable contents of the different data file types, and coding schemes, as well as documentation about the achievement items and context questionnaire items. A supplement to the user guide provides the National Adaptations documentation, describing national adaptions made to the PIRLS 2021 Context Questionnaires by individual countries and how the data were handled.