PIRLS case study analysis with SPSS

Mateusz Kleczaj (m.kleczaj@ibe.edu.pl)

Table of contents

1 Introduction

This guide shows how to use SPSS to work with data from international large-scale assessments. It is built around examples from **PIRLS 2021** (*Progress in International Reading Literacy Study*), which measures the reading comprehension of fourth-grade students. Studies of this kind are run by two main organisations: **IEA** (*International Association for the Evaluation of Educational Achievement*), responsible for TIMSS, PIRLS, ICCS and ICILS, among others, and **OECD**, which runs studies such as PISA, TALIS and PIAAC.

The guide is intended for those who want to get to know the data efficiently, prepare variables for analysis and obtain first descriptive results. It does not require advanced statistical knowledge. Its aim is to show what can be done in SPSS conveniently and correctly at the data-preparation stage, and also to indicate the point at which it is worth turning to tools designed specifically for analysing IEA and OECD data — above all the IEA IDB Analyzer, which works directly with SPSS and lets you carry out these analyses without switching software.

---

2 Why is SPSS useful for working with these data?

PIRLS data, like data from other international large-scale assessments, are distributed in the `.sav` format, the native SPSS format. As soon as a file is opened, all variable and value labels are visible — one of the clear advantages of working in SPSS, which offers both a readable graphical interface and the ability to work with syntax. You can quickly review the structure of the dataset, check response distributions, merge files from different questionnaires and prepare variables for further analysis, without having to install additional packages or pre-process the file.

SPSS is particularly useful for:

- opening and reviewing labelled `.sav` files
- merging files from different questionnaires (student, school, teacher)
- appending national (country-specific) questions to the international database
- checking missing data and variable distributions
- recoding variables and building simple indices
- computing sums and means from several questionnaire items
- producing preliminary exploratory tables and charts
- preparing an output file for further analysis

SPSS works well as a first tool when working with these data — it lets you quickly get oriented in the structure of the dataset, assess data quality and prepare variables for the actual analyses. However, once you want to test whether differences between groups of students are statistically significant, the standard SPSS procedures have important limitations that follow from the specific nature of these data. Although OECD provides dedicated SPSS macros that allow correct computation, they are considerably more complex to use than dedicated tools or packages such as **IEA IDB Analyzer**, **Rrepest**, **repest** or **intsvy**. Guides to these tools are available on the IBE-PIB website: ibe.edu.pl/en/data/how-to-analyze-data-from-international-large-scale-assessments-ilsa.

When SPSS is not enough

Task	SPSS	Dedicated tool
Reading and reviewing files		
Merging questionnaire files		
Recoding variables		
Exploratory tables and charts		
Computing weighted percentages		
Point estimates of means from PVs		
Correct standard errors		
Testing differences between groups		
Full analysis of achievement scores (PVs)		

SPSS does not work alone — but the IDB Analyzer completes it

The limitations in the table above do not mean you have to abandon SPSS. The IEA IDB Analyzer is a free tool that — unlike the Rrepest, repest or intsvy packages (which require R or Stata) — works directly with SPSS:

- it operates on the same `.sav` files,
- it requires no programming knowledge (point-and-click interface),
- it generates ready-made syntax that you run in SPSS, and you get the results in the SPSS output window,
- it automatically accounts for replicate weights and plausible values (PVs) — exactly what the standard SPSS procedures lack.

In practice you prepare the data in SPSS (merging files, `RECODE`, `COMPUTE`, building indices) and hand the final population estimates over to the IDB Analyzer — without leaving the SPSS ecosystem. The IDB Analyzer is therefore a complement to SPSS, not a replacement for it.
A guide to the IDB Analyzer is available on the IBE-PIB website: ibe.edu.pl/en/data/how-to-analyze-data-from-international-large-scale-assessments-ilsa/idb-analyzer-eng

To use SPSS knowingly and safely, it is worth understanding where these limitations actually come from. The specific nature of international large-scale assessments rests on three methodological pillars that substantially change the approach to statistical analysis.

3 The specific nature of international assessment data

Data from studies such as PIRLS, TIMSS or PISA differ from typical survey datasets in three important respects. We discuss them briefly below to justify the choices that appear later in the guide: why we use `WEIGHT BY`, why results from the standard SPSS procedures require careful interpretation, and where the several scores per student come from.

3.1 Sampling weights: why we use `WEIGHT BY`

In PIRLS, schools are selected with probability proportional to their size, and then one or two classrooms take part in each sampled school. This way of sampling means that individual schools have different chances of being selected. Without applying weights, the results of analyses do not reflect the population values for fourth-grade students.

To address this, each student is assigned a sampling weight indicating how many students in the population that person represents. In PIRLS 2021 this is the `TOTWGT` variable (the total student weight). The weight is activated with the `WEIGHT BY TOTWGT` command or through the *Data* → *Weight Cases…* menu.

> ⚠ Use `WEIGHT BY` so that results are representative
>
> Using `WEIGHT BY TOTWGT` is essential for the computed means and percentages to correspond to population values. Without the weight, results may be distorted by the over-representation of large or small schools, uneven questionnaire response rates and a range of other factors that follow from the sampling design.

3.2 The complex sampling design: why SPSS reports different standard errors

Students in PIRLS are not sampled individually from the whole population. Schools are sampled first, then classrooms. Because students from the same school or classroom are usually more similar to one another than students from different schools, the standard SPSS procedures — which assume that observations are fully independent — underestimate standard errors, confidence intervals and p-values relative to the methods used in international assessments.

To estimate the uncertainty of estimates correctly, replicate weights are used: sets of additional weights that reflect modified versions of the sample. They make it possible to account for the uncertainty arising from the complex sampling design. The lack of built-in support for replicate weights is one of the main limitations of SPSS when working with international assessment data.

⚠ Standard errors and significance tests require dedicated tools

The standard SPSS procedures do not support analysis based on replicate weights, so the standard errors, confidence intervals and p-values they produce should, in published results, be replaced with values from tools dedicated to international assessments — most simply with the IEA IDB Analyzer, which works directly on `.sav` files and returns results in SPSS. In exploratory analyses, the SPSS results can serve as a rough reference point.

3.3 Plausible values: where the several outcome variables come from

In a traditional test, every student answers all the questions and receives a single score. In studies such as PIRLS, however, there are far too many items for a single student to answer them all within the available testing time — so each student sees only a randomly assigned subset of items. To estimate a student's score on the full scale nonetheless, statistical methods are used that, instead of a single number, produce several **plausible values** (PVs).

For reading achievement in PIRLS 2021 there are five variables: `ASRREA01`, `ASRREA02`, `ASRREA03`, `ASRREA04` and `ASRREA05`. In other studies the number of PVs may differ — PISA and PIAAC use ten.

⚠ For exploration only

The mean of the PVs can be useful for exploring the data, but it should not be the basis for reporting analyses such as regression, correlations or statistical comparisons of groups. For such analyses you should use tools that handle PVs in line with international-assessment methodology — the IDB Analyzer, Rrepest, repest or intsvy.

---

4 PIRLS 2021 data: downloading and organising files

Before starting work in SPSS, you need to download the appropriate files and organise them on disk. It is worth distinguishing two types of material from the outset. The first is the **public international databases** made available by IEA — they contain data from all countries participating in the study and are available free of charge in the IEA repository. The second is **national material**: each participating country may ask respondents additional questions that are not part of the international instrument. Some countries make the answers to such questions publicly available. IBE-PIB provides the Polish national questions as separate datasets that can be merged with the international database using shared identifiers.

4.1 Where to download the data

PIRLS 2021 data are available free of charge from the IEA repository: www.iea.nl/data-tools/repository/pirls

Download the files in SPSS format (`.sav`). Together with the data, download the technical documentation, especially the *User Guide* and the *codebook*, which describe the variable names, response codes and identifiers needed to merge files.

If you are interested in the Polish data supplemented with national questions, you will find them in the IBE-PIB repository: www.ibe.edu.pl/en/data/where-to-find-the-data-and-research-instruments/pirls-data-international-assessments

4.2 Folder structure

Before you start, it is worth setting up a clear folder structure. This helps you keep control over the successive stages of the work and minimises the risk of accidentally overwriting the original data.

```
PIRLS2021/
   raw/        # original files downloaded from IEA, do not modify
   working/    # working files created in SPSS
   syntax/     # saved .sps syntax
   output/     # tables and charts
```

4.3 IEA file names

PIRLS files are split by topic — each questionnaire is a separate file, distinct for each country. Files are named following the pattern `[type][country][cycle].sav`. For Poland in PIRLS 2021:

File	Contents
`asgpolr5.sav`	Student questionnaire and achievement
`ashpolr5.sav`	Home (parent) questionnaire
`acgpolr5.sav`	School (principal) questionnaire
`atgpolr5.sav`	Teacher questionnaire
`asapolr5.sav`	Item-level responses
`astpolr5.sav`	Student–teacher linkage
`asppolr5.sav`	Process data / paradata: response times and process data (digitalPIRLS only)

The prefix indicates the data type: `asg` = student, `ash` = home (parent), `acg` = school, `atg` = teacher, `asa` = item responses, `ast` = student–teacher linkage, `asp` = paradata.

The country code `pol` is Poland. The cycle code `r5` is PIRLS 2021.

> **Where to start?**
>
> The `asgpolr5.sav` file contains both the student questionnaire data and the plausible values (PVs) as well as the `TOTWGT` weight — everything you need for basic analyses. This is the file to which we append the data from the remaining files.

5 Reading and reviewing data in SPSS

5.1 Opening a file and a first look at the data

PIRLS files are opened in the usual way through *File* → *Open* → *Data…* or with syntax:

```
GET FILE='C:\PIRLS2021\raw\asgpolr5.sav'.
```

Once opened, SPSS displays variable and value labels immediately in the *Data View* and *Variable View* tabs — with no extra steps required.

> **Save your syntax from the start**
>
> It is worth saving every operation in a syntax file (`.sps`). You open a syntax window through *File* → *New* → *Syntax*. You run it by selecting a block of code and clicking the green *Run* arrow.

It is good practice to inspect the structure of the file right after loading it. A quick way is to right-click the header of a column such as `IDCNTRY` and choose *Descriptive Statistics*.

If the file contains data for Poland only, the result will show a single value (616) together with the number of observations — in the Polish PIRLS 2021 sample this is 4,179 students. For a multi-country file you will see a list of all the countries.

5.2 Checking the identification variables

Every IEA file contains technical and identification variables used to merge data. The most important ones are listed below — not all of them appear in every file, but it is worth knowing them before merging datasets.

Variable	Description
`IDCNTRY`	Country code (616 for Poland)
`IDSCHOOL`	School identifier
`IDCLASS`	Class identifier
`IDSTUD`	Student identifier
`IDTEALIN`	Teacher identifier (in the `ast` and `atg` files)
`TOTWGT`	Total student weight
`JKZONE`	Jackknife zone (for estimating standard errors)
`JKREP`	Replicate indicator (0 or 1)

The `JKZONE` and `JKREP` variables are replication variables — as described in Section 3, we do not use them directly in the standard SPSS procedures. In dedicated analysis tools such as the IDB Analyzer or Rrepest, they are accounted for automatically.

---

6 Merging data files

PIRLS data are split into separate files for each questionnaire — as discussed in Section 4.3. To analyse, for example, the relationship between parental education and a student's achievement, you need to merge the student file with the home (parent) file. Merging is done **horizontally** — we add new variables to existing rows. The number of observations (students) stays the same; the set of variables grows.

Before merging files

Student, school and teacher identifiers are unique within a country, but not across the whole database — the same `IDSTUD` value can appear in the data of two different countries. That is why, in every merge, we always list `IDCNTRY` as the first key:

- **Student school:** `IDCNTRY` + `IDSCHOOL`
- **Student teacher (via AST):** `IDCNTRY` + `IDTEALIN`
- **Student home (parent) / national questions:** `IDCNTRY` + `IDSTUD`

6.1 Merging the student file with the home (parent) file

The merge procedure in SPSS requires both files to be sorted by the same key variables. In PIRLS, the keys for merging the student with the home (parent) file are `IDCNTRY` and `IDSTUD` — each parent questionnaire is assigned to a specific student. The student file is `asgpolr5.sav` and the home file is `ashpolr5.sav`.

Step 1: sort the student file and save it:

```
GET FILE='C:\PIRLS2021\raw\asgpolr5.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\student_sorted.sav'.
```

Step 2: sort the home (parent) file and save it:

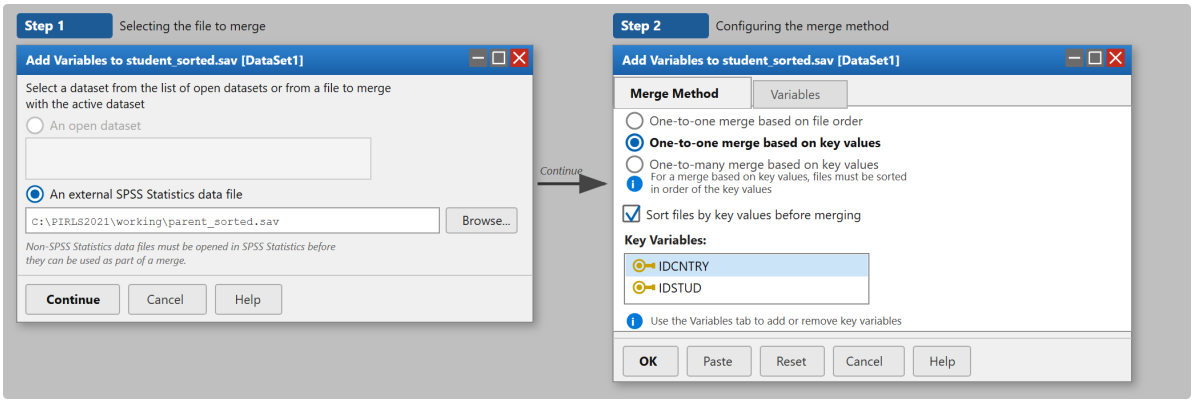
```
GET FILE='C:\PIRLS2021\raw\ashpolr5.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\parent_sorted.sav'.
```

Step 3: merge the files:

Through the menu: open `student_sorted.sav` as the active dataset, then: *Data → Merge Files → Add Variables…*

In the dialog, point to `parent_sorted.sav` as the external file. In the *Key Variables* field add `IDCNTRY` and `IDSTUD`. Tick the option *Match cases on key variables in sorted files*. Click *OK*.

```
MATCH FILES
  /FILE='C:\PIRLS2021\working\student_sorted.sav'
  /TABLE='C:\PIRLS2021\working\parent_sorted.sav'
  /BY IDCNTRY IDSTUD.
EXECUTE.
SAVE OUTFILE='C:\PIRLS2021\working\student_parent.sav'.
```



The resulting dataset — the student data supplemented with the parents' answers — will serve in the rest of the guide as the starting point for the worked examples in Section 8.

> **Check the result after merging**
>
> After merging, check that the number of rows is the same as before the merge. If it has increased, this signals a problem with the merge keys, most often a mismatch in variable types (numeric vs. string). The home file (`ash`) contains one row per student, so after a correct merge the number of observations should stay unchanged (for Poland: 4,179 students).

6.2 Appending teacher data

The student file (`asg`) and the teacher file (`atg`) have no shared identifier — they cannot be merged directly. The linkage file `ast` acts as the bridge: it contains both student and teacher identifiers. In Poland, each student has a single language-of-instruction teacher, so `ast` has the same number of rows as the student file and the merge proceeds just as with the other files.

We start from `asg` and append `ast` on the `IDCNTRY` + `IDSTUD` key — this brings `IDTEALIN` into the dataset. Then we append `atg` to this file on the `IDCNTRY` + `IDTEALIN` key.

```
* Step 1: append the teacher identifier to the student file.
GET FILE='C:\PIRLS2021\raw\asgpolr5.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\asg_sorted.sav'.

GET FILE='C:\PIRLS2021\raw\astpolr5.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\ast_sorted.sav'.

MATCH FILES /FILE='C:\PIRLS2021\working\asg_sorted.sav'
  /TABLE='C:\PIRLS2021\working\ast_sorted.sav'
  /BY IDCNTRY IDSTUD.
EXECUTE.
SAVE OUTFILE='C:\PIRLS2021\working\asg_ast.sav'.

* Step 2: append the teacher questionnaire data.
GET FILE='C:\PIRLS2021\working\asg_ast.sav'.
SORT CASES BY IDCNTRY IDTEALIN.
SAVE OUTFILE='C:\PIRLS2021\working\asg_ast_sorted.sav'.

GET FILE='C:\PIRLS2021\raw\atgpolr5.sav'.
SORT CASES BY IDCNTRY IDTEALIN.
SAVE OUTFILE='C:\PIRLS2021\working\atg_sorted.sav'.

MATCH FILES /FILE='C:\PIRLS2021\working\asg_ast_sorted.sav'
  /TABLE='C:\PIRLS2021\working\atg_sorted.sav'
  /BY IDCNTRY IDTEALIN.
EXECUTE.
SAVE OUTFILE='C:\PIRLS2021\working\student_teacher.sav'.
```

Merging teacher data is easier in the IDB Analyzer

If you plan analyses that use teacher variables, the IDB Analyzer will merge all the files automatically — without manual sorting or creating intermediate files. Manual merging in SPSS mainly makes sense when you want to stay in a single tool throughout the data-preparation stage.

6.3 Appending national questions

Each country participating in PIRLS may ask respondents additional questions that are not part of the international instrument. IBE-PIB provides the Polish national questions

as a separate dataset, which we merge with the international database on the IDCNTRY and IDSTUD keys: www.ibe.edu.pl/en/data/where-to-find-the-data-and-research-instruments/pirls-data-international-assessments

```
GET FILE='C:\PIRLS2021\working\student_parent.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\main_sorted.sav'.

GET FILE='C:\PIRLS2021\raw\PIRLS2021_PL_National.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\national_sorted.sav'.

MATCH FILES
  /FILE='C:\PIRLS2021\working\main_sorted.sav'
  /TABLE='C:\PIRLS2021\working\national_sorted.sav'
  /BY IDCNTRY IDSTUD.
EXECUTE.
SAVE OUTFILE='C:\PIRLS2021\working\pirls_poland_full.sav'.
```

---

7 Preparing variables

Before running the actual analyses, we need to prepare two working variables. In this section we activate the weight, show an example of recoding a variable from the Polish national database and compute the reading score `READ_MEAN`. The remaining variables used in Section 8 — `ASBH05AA` and `ASBGSLR` — are available directly in the `student_parent.sav` file and need no additional preparation.

7.1 Checking for missing data

In PIRLS files, missing data are usually already declared in the `.sav` format and visible in the *Missing* column of the *Variable View* tab. If, despite this, a value denoting a non-response is shown as a valid category, you can declare it manually:

```
MISSING VALUES ASBG05A (9).
```

7.2 Activating the sampling weight

Before starting analyses, you should activate the total student weight `TOTWGT` so that the results obtained correspond to population values.

Through the menu: click *Data* → *Weight Cases…*, tick the option *Weight cases by* and move the `TOTWGT` variable into the *Frequency Variable* field. Click *OK*. In the bottom-right corner of the SPSS window the label **Weight On** will appear, confirming that the weight is active.

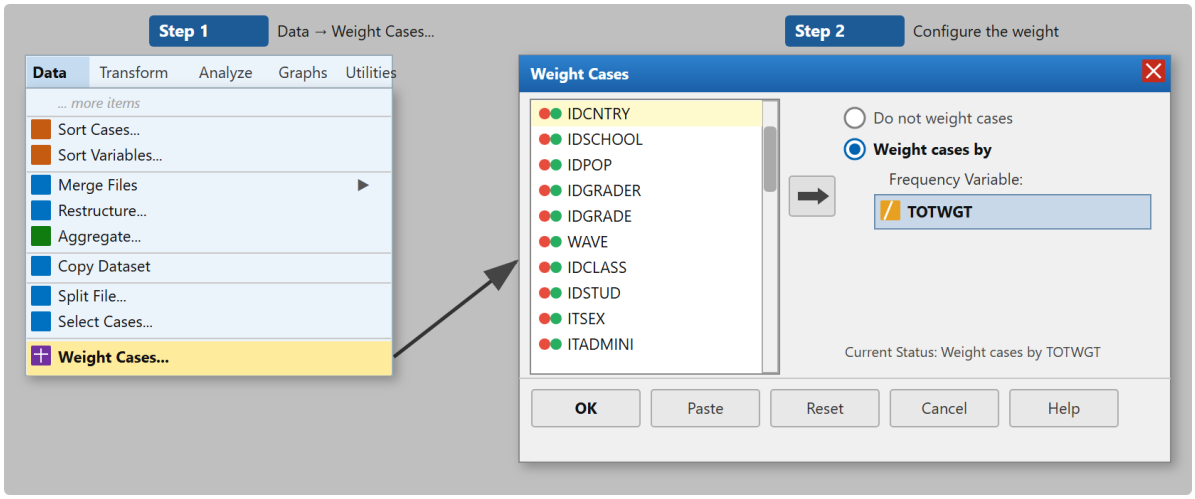


Figure 1: Weight Cases dialog: tick Weight cases by and move TOTWGT across

```
* Activate the sampling weight.
WEIGHT BY TOTWGT.

* ... your analyses here ...

* Turn the weight off when finished.
WEIGHT OFF.
```

> ⚠ The weight stays active until you turn it off
>
> After you run `WEIGHT BY`, SPSS applies the weight in all subsequent procedures until the file is closed or `WEIGHT OFF` is run. Always end a block of weighted analyses with the `WEIGHT OFF` command.

> ℹ Technical note: TOTWGT and the sample size
>
> `TOTWGT` sums to the population size (about 330,000 students in Poland), which makes SPSS treat a sample of 4,000 people as if it were a census of the whole population and produces drastically inflated F statistics and p-values. If you use the standard significance tests in SPSS, use the **`HOUWGT`** weight instead — it gives the same means and percentages but preserves the actual sample size, so the test results are less misleading. This does not mean, however, that `HOUWGT` is the correct weight for reporting analyses — it serves only to reduce the artefacts arising from the standard SPSS procedures.

7.3 Recoding variables

Variables from the PIRLS databases often need to be transformed before analysis — for example, by collapsing several categories into one or creating a binary (0/1) variable. The IDB Analyzer does not allow you to create new variables or recode — operations of this kind must be done beforehand in SPSS or another environment, and only then is the finished file loaded into the IDB Analyzer. SPSS is particularly useful here thanks to its intuitive interface and direct support for `.sav` files.

The example below shows recoding the variable `ASXH05A` from the Polish national database (`PIRLS2021_PL_National.sav`). The variable has six categories:

- 1 = Did not attend
- 2 = Less than one year
- 3 = One year
- 4 = Two years or more
- 9 = Omitted or invalid response
- 98 = Not administered

We recode it into a binary variable — attended nursery (1) or not (0):

Through the menu: *Transform → Recode into Different Variables…* Set `ASXH05A` as the input variable, give the new variable the name `NURSERY_BIN` and define the value mapping in the *Old and New Values* window.

```
RECODE ASXH05A
  (1=0) (2=1) (3=1) (4=1) (ELSE=SYSMIS)
  INTO NURSERY_BIN.
VARIABLE LABELS NURSERY_BIN 'Nursery attendance'.
VALUE LABELS NURSERY_BIN 1 'Yes' 0 'No'.
EXECUTE.
```

You can verify the result of the recode immediately by comparing the distributions of the two variables:

```
FREQUENCIES VARIABLES=ASXH05A NURSERY_BIN.
```

> **It is worth reaching for the national data**
>
> The counterpart of `ASXH05A` in the international database is **`ASBH05AA`**, which is already stored in dichotomous form (yes/no) and does not allow you to distinguish the duration of care. The Polish national question was expanded: instead of a single "yes" category, it distinguishes three variants by length of attendance, which allows a more detailed analysis — for example, examining whether the *length* of nursery attendance is associated with reading achievement.
> When reviewing the documentation of the national questions, it is worth checking each time whether the Polish data offer a richer measurement than the international version.

7.4 Computing the READ_MEAN variable (mean of the PVs)

We compute the outcome variable `READ_MEAN` as the arithmetic mean of the five plausible values (`ASRREA01`–`ASRREA05`).

Through the menu: *Transform → Compute Variable…* In the *Target Variable* field enter `READ_MEAN`. In the *Numeric Expression* field enter `MEAN(ASRREA01, ASRREA02, ASRREA03, ASRREA04, ASRREA05)`. Click *OK*.

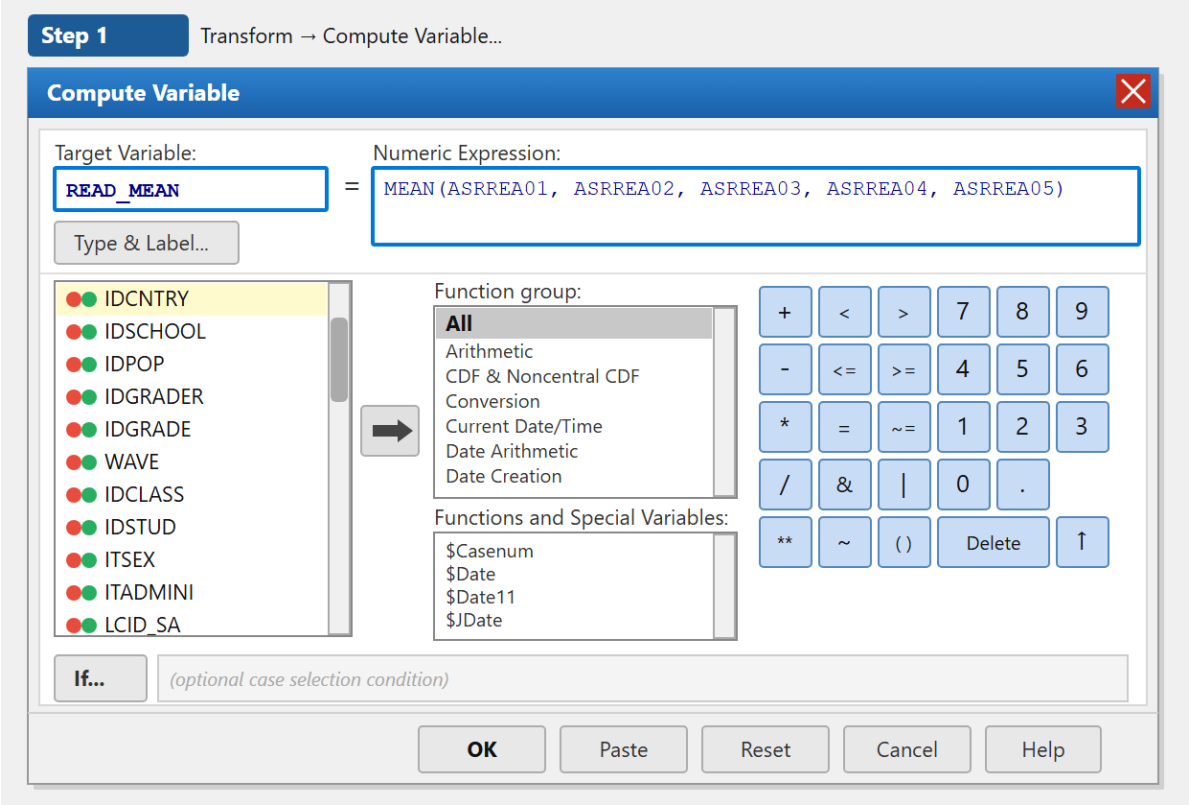


Figure 2: Compute Variable window: computing the mean of the five plausible values

```
COMPUTE READ_MEAN = MEAN(ASRREA01, ASRREA02, ASRREA03, ASRREA04, ASRREA05).
VARIABLE LABELS READ_MEAN 'Approximate reading score (mean of 5 PVs)'.
EXECUTE.
```

An approximate student score

Plausible values are not designed as exact scores for individual students — they are a tool for making inferences about the population, not about the individual. `READ_MEAN` is therefore, by design, an approximation of a particular student's score: it reflects their likely level of proficiency, not a value "measured" with full precision.
For simple group means, the limitation is not the point estimate itself but the standard errors and significance tests.

8 Worked examples, step by step

The section below presents a few typical analyses based on the PIRLS 2021 data for Poland. Each example includes: the aim of the analysis, the path through the SPSS graphical menu, the corresponding syntax and a note on interpreting the results. We use the `student_parent.sav` file (student merged with parent, Section 6.1) and the computed `READ_MEAN` variable, as described in the previous section. In Example 8.2 we additionally use the `NURSERY_BIN` variable recoded from the national question `ASXH05A` in Section 7.3.

8.1 Reading achievement by gender

Aim of the analysis: Compare the reading achievement of girls and boys. The means match the figures presented in Table 5.5 of the PIRLS 2021 national report (p. 65, in Polish) and are reproduced in this example.

The dependent variable is `READ_MEAN` — the reading score computed as the mean of the five PVs. The independent variable is `ITSEX`, the student's gender: `1 = girl`, `2 = boy`.

Through the menu: *Analyze* → *Compare Means and Proportions* → *Means…* Move `READ_MEAN` into the *Dependent List* field and `ITSEX` into the *Independent List* field. Click *OK*.

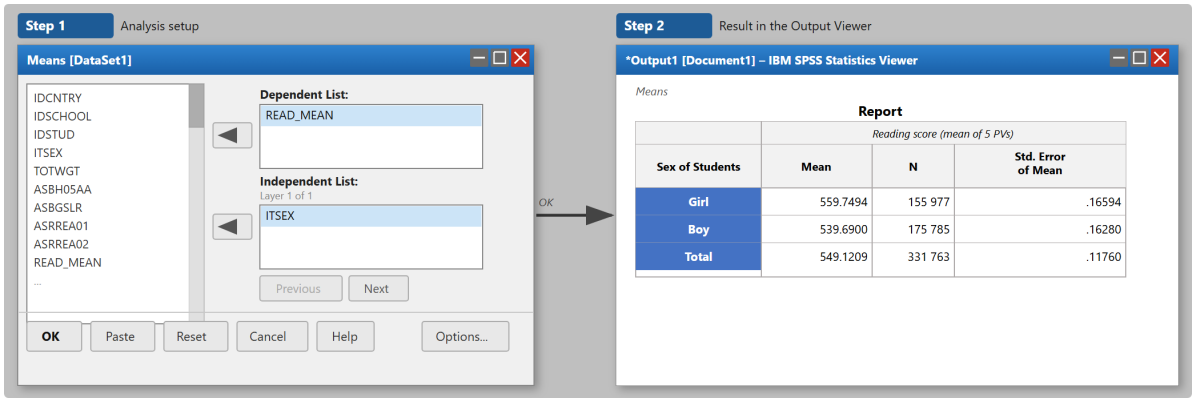


Figure 3: Means dialog and procedure output: reading achievement by student gender

```
WEIGHT BY TOTWGT.

MEANS TABLES=READ_MEAN BY ITSEX
  /CELLS=MEAN COUNT SEMEAN.

WEIGHT OFF.
```

How to read the results? The `MEANS` procedure returns a table with columns: mean, `N` and the standard error of the mean. The mean shows the reading score in a given group, and `N` corresponds to the weighted number of students in the population, not the sample size.

The group means obtained with this method match the national report: girls score about 559.75 points and boys about 539.69 points. The difference is therefore about 20 points in favour of girls. This means that, in the PIRLS 2021 data, girls achieved on average higher reading scores than boys.

> ⚠ Standard errors from SPSS are underestimated
>
> The standard errors shown in the SPSS table are much lower than the correct values from the national report. This happens because the standard `MEANS` procedure does not account for replicate weights and the complex sampling design. In addition, using `TOTWGT` makes SPSS treat the weighted sample as the population, which further inflates the precision of the estimates.
> The means can be treated as correct point estimates and used for exploring the data. To assess the significance of the difference between girls and boys, however, you should use dedicated tools such as the IEA IDB Analyzer, Rrepest, repest or intsvy.

8.2 Reading achievement and nursery attendance

Aim of the analysis: Check whether students who attended nursery achieve higher reading scores. As the independent variable we use `NURSERY_BIN` (1 = Yes, 0 = No) — the variable recoded in Section 7.3 from the national question `ASXH05A`.

Because `NURSERY_BIN` comes from the Polish national database, not from the `student_parent.sav` file, before the analysis we need to append the national database to our working file. The procedure is analogous to the one described in Section 6.3: we open `student_parent.sav`, sort it by `IDCNTRY` and `IDSTUD`, and then append the variables from the `PIRLS2021_PL_National.sav` file sorted by the same keys.

```
GET FILE='C:\PIRLS2021\working\student_parent.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\student_parent_sorted.sav'.

GET FILE='C:\PIRLS2021\raw\PIRLS2021_PL_National.sav'.
SORT CASES BY IDCNTRY IDSTUD.
SAVE OUTFILE='C:\PIRLS2021\working\national_sorted.sav'.

MATCH FILES
  /FILE='C:\PIRLS2021\working\student_parent_sorted.sav'
```

```
  /TABLE='C:\PIRLS2021\working\national_sorted.sav'
  /BY IDCNTRY IDSTUD.
EXECUTE.
SAVE OUTFILE='C:\PIRLS2021\working\student_parent_pl.sav'.
```

The dependent variable is `READ_MEAN`, the independent variable is `NURSERY_BIN`.

Through the menu: *Analyze* → *Compare Means and Proportions* → *Means…* Move `READ_MEAN` into the *Dependent List* field and `NURSERY_BIN` into the *Independent List* field. Click *OK*.

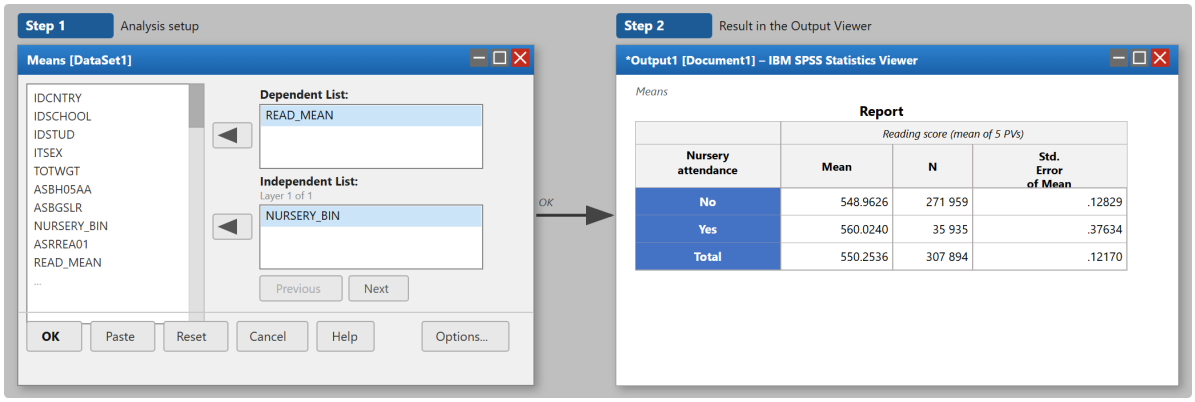


Figure 4: Means dialog and procedure output: reading achievement and nursery attendance

```
WEIGHT BY TOTWGT.

MEANS TABLES=READ_MEAN BY NURSERY_BIN
  /CELLS=MEAN COUNT SEMEAN.

WEIGHT OFF.
```

How to read the results? The output table will have two rows: one for students who attended nursery (`NURSERY_BIN = 1`) and one for those who did not (`NURSERY_BIN = 0`). The *Mean* column shows the reading score in each group.

Students who attended nursery achieve, on average, higher reading scores. The difference should be interpreted with caution: nursery attendance is strongly associated with the family's socio-economic status and other variables, so the raw difference in means need not reflect a causal effect of nursery care.

8.3 Students' attitudes to reading and gender — distribution of categories

Aim of the analysis: Compare the distribution of attitudes to reading between girls and boys. The variable `ASDGSLR` is a categorical index (*Students Like Reading Index*) taking three values: 1 = Very Much Like Reading, 2 = Somewhat Like Reading, 3 = Do Not Like Reading. The independent variable is `ITSEX` (student's gender).

The analysis uses the `student_parent.sav` dataset prepared in Section 6.1.

Through the menu: *Analyze* → *Descriptive Statistics* → *Crosstabs…* Move `ITSEX` into the *Row(s)* field and `ASDGSLR` into the *Column(s)* field. Click *Cells…*, tick *Row* in the *Percentages* section, then *Continue* and *OK*.

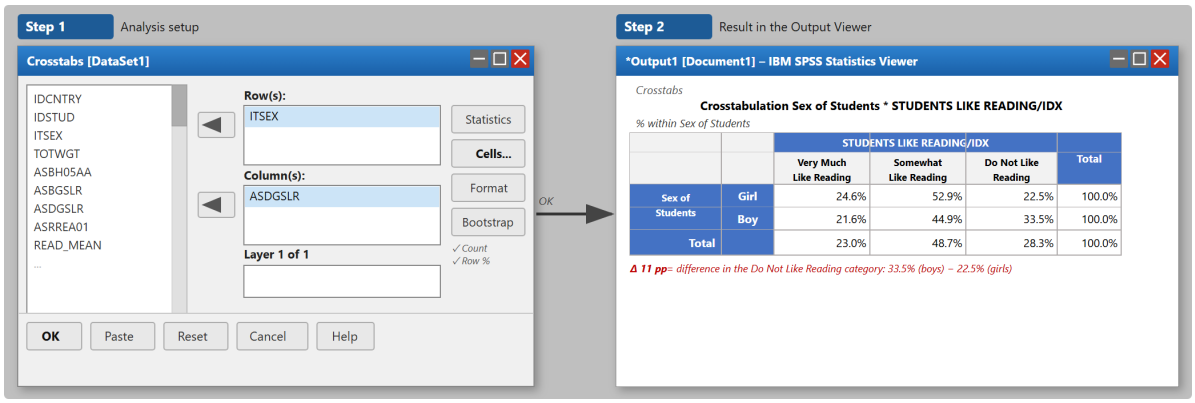


Figure 5: Crosstabs dialog and output: attitudes to reading by student gender

```
WEIGHT BY TOTWGT.

CROSSTABS
  /TABLES=ITSEX BY ASDGSLR
  /CELLS=COUNT ROW.

WEIGHT OFF.
```

How to read the results? The percentages in the table sum to 100% in each row (% of Sex of Students), which lets you compare the distribution of attitudes to reading between girls and boys regardless of the group sizes.

The results reveal a clear difference: in the *Do Not Like Reading* category there were 22.5% of girls and as many as 33.5% of boys — a difference of 11 percentage points. Boys also less often declared that they very much like reading (21.6% vs 24.6%). Such large differences in attitudes to reading are consistent with the differences in test scores described in Section 8.1 and provide important interpretive context for these data.

9 Good practices for working with data in SPSS

- **Do not modify the original files.** Keep the raw data in a separate folder and never overwrite it. Carry out all operations on working copies.
- **Save your syntax.** Tables and charts generated through the menu leave no trace. Saved syntax lets you reproduce every step.
- **Check the variable documentation.** Variable names change between study cycles. Before each analysis, make sure you are using the current names.
- **Keep an eye on missing data.** Before using a variable, check the range of values and the missing codes in *Variable View* or with `FREQUENCIES`.
- **Apply the weight deliberately.** Use `WEIGHT BY` before an analysis and `WEIGHT OFF` after it. Check that the status bar shows *Weight On* before running a procedure.
- ***p*-values and standard errors from SPSS are for orientation only.** In exploratory analyses they can be treated as a signal of whether a phenomenon is worth investigating more deeply. In published results you should use values from the IDB Analyzer or Rrepest.
- **Verify your results.** Compare the percentages and means you obtain with the official national reports to make sure the data have been read in correctly.

9.1 Most common mistakes

The problems below come up regularly when working with international assessment data in SPSS. It is worth knowing about them before they occur.

The number of rows increased after merging files. This signals an error in the merge keys. The most common cause is a mismatch in the types of the key variables (e.g. numeric in one file and string in the other) or omitting `IDCNTRY` in the merge — SPSS then matches students from different countries who share the same `IDSTUD`. Solution: check the variable types in *Variable View* and make sure both files are sorted by the same keys before `MATCH FILES`.

The weight was left on from a previous analysis. If `WEIGHT OFF` was omitted, every subsequent procedure still uses the weight. Check the SPSS status bar — the *Weight On* label means the weight is active. As a matter of good practice, every block of analysis should end with the `WEIGHT OFF` command.

`TOTWGT` gives a huge N in the table. This is expected behaviour — the weight scales the results up to the population level (about 330,000 fourth-graders for Poland). The *N* column in a weighted table is not the sample size. To assess the actual sample size in exploratory analyses, you can use an unweighted `FREQUENCIES`.

A variable has missing values coded as numeric categories. Values such as 9, 99 or 998 may denote "no response", but SPSS treats them as data if they have not been declared as

missing. This leads to inflated or incorrect results. Before each analysis, check the codebook and declare the missing values in *Variable View* → *Missing*.

The files are not sorted before merging. `MATCH FILES` with the `BY` option requires both files to be sorted by the key variables in the same order. Failing to sort produces incorrect matches with no error message at all.

The `asg` file was confused with `asa` or another. In PIRLS, `asg` is the student file, `ash` the home (parent) file, `acg` the school file and `asa` the file with item-level responses. Loading the wrong file often produces no error — variables are simply missing or the data look odd. Always verify the contents of a newly opened file through *Data* → *Variable View*.

The standard error values do not match the report. Differences on the order of 10–20 times between the SE from SPSS and the SE from the national or international report are a normal effect of the lack of support for replicate weights and the inflated precision when using `TOTWGT` (see Section 3.2). In published results you should always report SEs computed with dedicated tools.

---

10 Summary

SPSS is a useful tool at the stage of preparing and initially exploring data from international large-scale assessments. Its intuitive graphical interface and native support for the `.sav` format mean that it works well for reading in data, merging files, appending national questions, checking variable distributions, recoding and creating simple visualisations.

On its own, however, SPSS does not handle replicate weights or plausible values, so for reporting analyses — correct standard errors, significance tests and the full analysis of achievement scores — it is not sufficient. This does not mean you have to give it up. The simplest, SPSS-compatible complement is the IEA IDB Analyzer: it works on the same `.sav` files, requires no programming and returns results in the SPSS output window, while correctly accounting for replicate weights and PVs.

The recommended approach is therefore to prepare and explore the data in SPSS, and then to carry out the actual population analyses in the IDB Analyzer (or — if you work in R or Stata — in the Rrepest, repest or intsvy packages). This two-stage way of working lets you take advantage of the convenience of SPSS at the data-preparation stage while preserving the methodological correctness of the final results. It is good practice to verify descriptive results by comparing them with the official national reports or with results from the IDB Analyzer, which confirms that the data have been read in and prepared correctly.

11 Resources and references

11.1 Where to download data

- **PIRLS and TIMSS**: www.iea.nl/data-tools/repository
- **ICCS and ICILS**: www.iea.nl/data-tools/repository
- **PISA**: www.oecd.org/pisa/data
- **TALIS**: www.oecd.org/en/about/programmes/talis.html#data
- **PIAAC**: www.oecd.org/skills/piaac/data
- **SSES** (*Survey on Social and Emotional Skills*): www.oecd.org/en/about/programmes/oecd-survey-on-social-and-emotional-skills.html#data
- **Polish national questions and IBE-PIB data**: www.ibe.edu.pl/en/data

11.2 Technical documentation

- Fishbein, B., Yin, L., & Foy, P. (2024). *PIRLS 2021 User Guide for the International Database* (2nd ed.). Boston College. pirls2021.org/data
- Fishbein, B., Taneva, M., & Kowolik, K. (2025). *TIMSS 2023 User Guide for the International Database*. Boston College. timss2023.org/data
- OECD (2023). *PISA 2022 Technical Report*. OECD Publishing. www.oecd.org/pisa/data/2022database

11.3 Dedicated tools for analyses with PVs and replicate weights

For the correct estimation of standard errors, testing differences between groups and the full analysis of achievement scores with plausible values, tools designed specifically for working with IEA and OECD data are essential. For each of them, IBE-PIB has prepared a dedicated guide available at the link below.

- **IEA IDB Analyzer** – a graphical, windowed program that generates SPSS, SAS or R syntax; it supports all IEA studies as well as PISA. It requires no programming knowledge.
  Download: www.iea.nl/data-tools/tools
  IBE-PIB guide: ibe.edu.pl/en/data/how-to-analyze-data-from-international-large-scale-assessments-ilsa/idb-analyzer-eng
- **Rrepest** – an R package developed by OECD; it supports PISA, PIAAC, TALIS and SSES. It allows flexible analyses of regressions, means and percentiles with full support for PVs and replicate weights.
  CRAN: cran.r-project.org/package=Rrepest
  IBE-PIB guide: ibe.edu.pl/en/data/how-to-analyze-data-from-international-large-scale-assessments-ilsa/r-repest-eng

- **intsvy** – an R package supporting a wide range of studies (PIRLS, TIMSS, PISA, PIAAC, ICILS and others); it allows the computation of means, percentages, correlations and regression models with replicate weights.
  CRAN: cran.r-project.org/package=intsvy
  IBE-PIB guide: ibe.edu.pl/en/data/how-to-analyze-data-from-international-large-scale-assessments-ilsa/intsvy-eng
- **repest** – a command for Stata developed by OECD; it supports PISA, PIAAC and TALIS. Avvisati, F., & Keslair, F. (2014).
  IBE-PIB guide: ibe.edu.pl/en/data/how-to-analyze-data-from-international-large-scale-assessments-ilsa/stata-repest-eng