User Guide for the German PIAAC Scientific Use File: Version II

Perry, Anja; Helmschrott, Susanne; Konradt, Ingo; Maehler, Débora B.

gesis Leibniz-Institut für Sozialwissenschaften

GESIS *Papers* 2017|23

User Guide for the German PIAAC Scientific Use File

Version II

Anja Perry, Susanne Helmschrott,
Ingo Konradt & Débora B. Maehler

GESIS Papers 2017|23

User Guide for the German PIAAC Scientific Use File

Version II

*Anja Perry, Susanne Helmschrott, Ingo Konradt &
Débora B. Maehler*

GESIS – Leibniz-Institut für Sozialwissenschaften 2017

GESIS Papers

GESIS – Leibniz Institute for the Social Sciences
Research Data Center PIAAC (FDZ PIAAC)
PO Box 12 21 55
68072 Mannheim
Germany

E-Mail: fdz-piaac@gesis.org
www.gesis.org/en/piaac/rdc

ISSN: 2364-3781 (Online)
Publisher, printing
and distribution: GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Cologne, Germany

Content

This User Guide gives a brief overview of the German PIAAC Scientific Use File [ZA5845][1] and information necessary for doing basic analyses using the PIAAC data.

Further information on PIAAC can be found on the following homepages
https://www.oecd.org/skills/piaac/data (OECD),
https://www.gesis.org/en/piaac/rdc (GESIS)

as well as in the PIAAC Technical Reports by Zabal et al. (2014) and the OECD (2013b, 2019).

1 Getting started

Getting the German PIAAC Scientific Use File

The German PIAAC Scientific Use File (Study No. ZA5845)[2] is released for academic research only. You can receive the current version of the data set (Rammstedt et al., 2016) after filling out and signing a data distribution contract. The contract is available here:

https://www.gesis.org/en/piaac/rdc/data/national-scientific-use-files

Please pay attention to the terms of use (usage- and charge regulations) as well as the information on proper citations specified in the contract. The signed contract must be sent to:

GESIS – Leibniz Institute for the Social Sciences
Research Data Center PIAAC (FDZ PIAAC)
PO Box 12 21 55
68072 Mannheim
Germany

Once GESIS received a signed contract the dataset will be made available in SPSS and Stata format. Additional documents such as background questionnaire and code book are available at Research Data Center PIAAC webpage[3].

Accessing the PIAAC Public Use File(s)

The international PIAAC Public Use Files (PUF) contain data sets of more than 30 countries that participated in PIAAC until today (Round 1, Round 2, and Round 3). The Public Use Files including Germany can be downloaded for free at:

https://webfs.oecd.org/piaac/puf-data

The Public Use File(s) are available in SPSS and SAS format as well as a CSV-file.

The Cyprus PIAAC Public Use File is also freely available at:

https://doi.org/10.4232/1.12632

[1] Study No. by GESIS Data Archive, Cologne

[2] Title: Programme for the International Assessment of Adult Competencies (PIAAC), Germany - Reduced Version

[3] https://www.gesis.org/en/piaac/rdc

In the following you can find a list of PIAAC participating countries with their country name (country code) which are available as PIAAC Public Use Files.

Round-1 countries:

Australia (36), Austria (40), Belgium (56), Canada (124), Cyprus (196), Czech Republic (203), Denmark (208), Estonia (233), Finland (246), France (250), Germany (276), Ireland (372), Italy (380), Japan (392), Korea (410), Netherlands (528), Norway (578), Poland (616), Russian Federation (643), Slovak Republic (703), Spain (724), Sweden (752), United Kingdom (826), United States (840)

Round-2 countries:

Chile (152), Greece (300), Indonesia (360), Israel (376), Lithuania (440), New Zealand (554), Singapore (702), Slovenia (705), Turkey (792)

Round-3 countries:

Ecuador (218), Hungary (348), Kazakhstan (398), Mexico (484), Peru (604), United States (840)

Due to national data privacy legislation, the German Public Use File offers only restricted access to the German data. For example, variables such as age were coarsened for the Public Use File in order to prevent the re-identification of participants in PIAAC. The German Scientific Use File provides more detailed data.

Merging the PIAAC Scientific Use File with the Public Use File(s)

The German PIAAC Scientific Use File can be merged with the Public Use File in order to perform cross-country analyses. When merging the data sets for the different countries, the variables SEQID and CNTRYID_E should be used as identifiers. While SEQID is a unique identification key within each country data set, it is not unique across countries. Thus, an identifier combining both variables needs to be created.

Variable labels are identical throughout all Public Use Files. Usually, variable labels in the German Scientific Use File are identical to the Public Use File. However, labels in the German Scientific Use File differ in variables that include country information when categories are collapsed for data confidentiality reasons (e.g., CNT_CITSHIP). Label information may get lost when data sets are merged. Therefore we recommend the following SPSS syntax when merging the German PIAAC Scientific Use File with the PIAAC Public Use File:

```
* Create identifier in each dataset

GET FILE = "ZA5845_v2-2-0.sav".
COMPUTE IDENT = CNTRYID_E*100000 + SEQID.
SAVE OUTFILE = "ZA5845_v2-2-0.sav".

GET FILE = "prgautp1.sav".
COMPUTE IDENT = CNTRYID_E*100000 + SEQID.
SAVE OUTFILE = "prgautp1.sav".

* Repeat step for additional countries

* Sort data by identifier

GET FILE = "ZA5845_v2-2-0.sav".
SORT CASES by IDENT.
SAVE OUTFILE = "ZA5845_v2-2-0.sav".

GET FILE = "prgautp1.sav".
SORT CASES by IDENT.
SAVE OUTFILE = "prgautp1.sav".

* Repeat step for additional countries

* Merge of data files:

MATCH FILES
   /FILE = "ZA5845_v2-2-0.sav"
   /FILE = "prgautp1.sav"
   /BY IDENT.

SAVE OUTFILE = "DEU_AUT_merged.sav".
```

It is important for the German Scientific Use File to appear first in the MATCH FILES command, in order to keep additional label information for the German data on countries.

The IDB-Analyzer (see "Analyses using SPSS, SAS and Stata" below) in combination with SPSS can also be used to merge the German PIAAC Scientific Use File with the Public Use Files of other participating countries. When using the merge module of the IDB-Analyzer, please note that the names of all data files have to follow this convention:

prgXXXp1.sav

XXX needs to be replaced by the three letter code of the respective country. Thus, before merging the German Scientific Use File with data files of other countries using the IDB analyzer, the German data file needs to be renamed to prgdeup1.sav.

Also when merging the files using the IDB-Analyzer, it is recommended that the German data file appears first in the mask of the merge module in IDB. This is to ensure that additional label information for the German data on countries is included in the final dataset.

2 Important (additional) variables

The data file contains competency scores (*plausible values*) for each participant in the domains

- Literacy
- Numeracy
- Problem solving in technology-rich environments

as well as their background information.

Compared to the Public Use File available at the OECD homepage, the German Scientific Use File includes additional variables and more detailed variables. Additional data included in the Scientific Use File are, for example:

- German federal states
- Municipality size, coarsened
- Age, continuous
- Income, continuous
- Income, as reported in the background questionnaire
- Educational background, by categories of the German educational system
- More detailed information on origin
- More detailed information on languages spoken
- Profession, ISCO 4-digit
- Industry, ISIC 4-digit
- Time of day and weekday of interview
- Additional weighting information

3 Weighting

In the German Scientific Use File, the final PIAAC weight SPFWT0 is provided, together with its 80 replicate weights (SPFWT1-SPFWT80). In case the available analysis tools offered for PIAAC are used, the final weight and its replicate weights are automatically taken into account.

The final weight was computed by subsequent weighting steps adjusting the sample data for bias resulting from survey errors such as sampling error, nonresponse error or noncoverage error. The use of the final PIAAC weight thus enables unbiased inferences from the sample data to the general population of the 16 to 65 year-old residents in Germany. For example, without the application of the final weights, the share of persons with a low level of education would be underestimated because they were less likely to participate in PIAAC.

The different weighting steps included a) base weights, b) an unknown eligibility adjustment, c) a non-literacy-related nonresponse adjustment, d) a literacy-related nonresponse adjustment and e) calibration of the sample data to match German Microcensus data. If necessary, large weights were trimmed and data was recalibrated.

Variables used for the unknown eligibility adjustment and the non-literacy-related nonresponse adjustment were age, citizenship and municipality size. In the calibration step, data was poststratified to match Microcensus population totals for age, gender, region and education.

All cases with a completed background questionnaire[4] and literacy-related non-respondents with information on age and gender available (see "LRNR" below) are part of the net sample and thus received a final weight.

Further information on the weighting adjustments and the selection of weighting variables can be found in the national PIAAC Technical Report Germany by Zabal et al. (2014) and the International PIAAC Technical Report (OECD, 2013b).

[4] In Germany, two respondents did not complete the background questionnaire all the way until the end. However, they answered a sufficient amount of questions which still qualified them to be included in the net sample.

4 Literacy-related non-respondents (LRNR)

Some adults in the participating countries were not able to participate or complete the background questionnaire due to language problems, reading and writing difficulties or learning and mental disabilities. In some of these cases, the interviewer was able to collect basic information for these people, i.e. gender and age. According to international requirements these adults, for whom basic information could be collected, are part of the net sample. They are regarded as a part of the PIAAC target population that cannot be represented by survey respondents because they are supposedly different from respondents regarding their proficiency (OECD, 2013b; Rammstedt, 2013). These so-called literacy-related non-respondents (LRNR) account for 1.6 % of the German sample (see Table 1). This percentage varies across the participating countries. The variable QCFLAG_LR can be used to identify these cases. Category 1 of QCFLAG_LR represents the literacy-related non-respondents and category 0 all respondents.

Table 1: Quality control flag for 100% verification of literacy-related BQ NRs age and gender (QCFLAG_LR)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Not BQ literacy-related NR	5379	98.4	98.4	98.4
	1 Literacy-related BQ NR, age and gender successfully collected and values provided in AGE_LR and GENDER_LR	86	1.6	1.6	100
	Total	5465	100	100	

5 Competency information on population level

PIAAC data can only be used for estimating competencies on the level of populations or subgroups of populations. It does not allow estimating competencies of individuals.

Due to time constraints, every person responded to only a subset of the test items. The item set administered to the respondent depended on information provided in the background questionnaire, performance in solving previous item, and a random element (Zabal et al., 2014, p. 15-20). Because respondents answer only a subset of items, accuracy of individual assessment of competencies is considerably lower.

However, using multiple values, so-called plausible values, representing the distribution of a respondent's proficiency, accounts for the uncertainty resulting from measuring it with only a subset of the item pool. These values are unbiased estimates on the group-level. They are based on responses to the subset of items and on background information (for further information on Item-Response-Theory see Mislevy, 1991; Von Davier, Gonzalez & Mislevy, 2009).

6 Analyzing PIAAC data: Using plausible values and replicate weights

Ten plausible values (e.g., PVLIT1 to PVLIT10) for each individual were derived using item response theory (IRT). Additionally, in order to account for the complex sample design applied in PIAAC, replicate weights were assigned to each individual.

Both the IRT and replicate approach must be taken into account when analyzing the PIAAC data. Ignoring either one will underestimate the error variance and therefore also the standard error. This user guide will provide a brief overview on how to take plausible values and the replicate approach into account when analyzing PIAAC data. For further information on plausible values and replicate weights see Von Davier, Gonzalez & Mislevy (2009), Statistics Canada (2002) and various sources from the OECD (2009, 2013a, 2013b).

In a first step we look at the use of plausible values when analyzing PIAAC data. When competencies are estimated, the estimate ($\hat{t}$) has to be computed using all plausible values (PV) across all individuals. Then the results of these computations are averaged:

$$\hat{t} = \frac{1}{m} \sum_{PV=1}^{m} \hat{t}_{PV} \text{ , } m = \text{number of plausible values}$$

A regression on literacy is therefore performed 10 times, resulting in 10 different parameters for the intercept and each coefficient. The final regression equation results from averaging the ten parameters. In Table 2 the results for a fictitious example of a regression are presented.

Table 2: Fictitious results for individual regressions with 10 plausible values

Germany	PVLIT 1	PVLIT 2	PVLIT 3	PVLIT 4	PVLIT 5	PVLIT 6	PVLIT 7	PVLIT 8	PVLIT 9	PVLIT 10
Intercept	0.125	0.128	0.124	0.126	0.121	0.126	0.125	0.120	0.121	0.127
Beta 1	1.052	1.058	1.047	1.056	1.062	1.051	1.053	1.061	1.055	1.054
Beta 2	0.583	0.592	0.546	0.498	0.646	0.574	0.526	0.588	0.623	0.456
Beta 3	3.227	4.124	3.014	3.541	3.208	3.610	3.429	4.072	3.034	3.257

The final estimates are calculated by averaging the estimates of the individual regressions.

Overall intercept:

(0.125 + 0.128 + 0.124 + 0.126 + 0.121 + 0.126 + 0.125 + 0.120 + 0.121 + 0.127) / 10 = **0.124**

Overall $\boldsymbol{\beta_1}$:

(1.052 + 1.058 + 1.047 + 1.056 + 1.062 + 1.051 + 1.053 + 1.061 + 1.055 + 1.054) / 10 = **1.055**

Overall $\boldsymbol{\beta_2}$:

(0.583 + 0.592 + 0.546 + 0.498 + 0.646 + 0.574 + 0.526 + 0.588 + 0.623 + 0.456) / 10 = **0.563**

Overall $\boldsymbol{\beta_3}$:

(3.227 + 4.124 + 3.014 + 3.541 + 3.208 + 3.610 + 3.429 + 4.072 + 3.034 + 3.257) / 10 = **3.452**

Thus, for this fictitious example the final regression equation results in

$$Y = \mathbf{0.124} + \mathbf{1.055}\beta_1 + \mathbf{0.563}\beta_2 + \mathbf{3.452}\beta_3 + \varepsilon$$

In a second step, we additionally take the complex sample design into account when computing the error variance. The correct estimation of the error variance is crucial when analyzing large-scale assessment data, such as PIAAC. The error variance of statistics in PIAAC consists of two components:

- Sampling variance (always present)
- Imputation variance (when plausible values play a role)

In the data collection process in PIAAC a complex sample design was used. Using the replicate approach, subsamples were drawn from the PIAAC sample. Each subsample represents the full sample. For each of the subsamples replicate weights were created. After the creation of replicate base weights, all weighting adjustments that were conducted for the full sample were conducted for each replicate sample to capture the variation created or reduced by the weighting adjustments (variables SPFWT1-SPFWT80). Across countries, different approaches with different numbers of replicate weights were applied in PIAAC, depending on the sampling design. Variables VEMETHOD or VEMETHODN indicate which approach was used in each country. In Germany, the Delete-one jackknife (JK-1) approach with 80 replicate weights was used.

The replicate weights are taken into account when computing the sampling variance. When $\hat{t}_i$ is the parameter computed using replicate weight *i*, the sampling variance for JK-1 is:

$$Var_{smpl} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{t}_i - \hat{t})^2 \text{ , } n = \text{number of replicate samples}$$

Please note that a different formula is valid when another replicate approach than JK-1 is used (see OECD, 2013b, p. 15 for details).

When competencies do not play a role in the estimation, the variance of the estimate equals the sampling variance:

$$Var(\hat{t}) = Var_{smpl}(\hat{t})$$

When competencies are estimated, thus plausible values play a role, the imputation variance needs to be added to the sampling variance. The imputation variance is computed as follows:

$$Var_{imp} = \frac{1}{m-1} \sum_{PV=1}^{m} (\hat{t}_{PV} - \hat{t})^2 \text{ , } m = \text{number of plausible values}$$

The final variance is then:

$$Var(\hat{t}) = \left(\frac{1}{m}\right) Var_{smpl}(\hat{t}) + \left(1 + \frac{1}{m}\right) Var_{imp}(\hat{t}) \text{, } m = \text{number of plausible values}$$

The standard error is the square root of the error variance:

$$SE = \sqrt{Var(\hat{t})}.$$

In order to appropriately compute the error variance when plausible values are taken into account, with the German PIAAC data one needs to perform 81 x 10 = 810 computations. That is considering 80 replicate weights plus one final weight (SPFWT0) times 10 plausible values.

Tools for analyzing the PIAAC data, acknowledging plausible values and the replicate approach, are available for SPSS, SAS, Stata and R. Each tools offer different and limited sets of analysis procedures.

Analyses using SPSS, SAS and Stata

The OECD provides analysis tools that account for the replicate design and plausible values (if applicable). These tools as well as comprehensive documentation of the PIAAC data are available from the OECD homepage:

https://www.oecd.org/skills/piaac/data

For analyses using SPSS the IDB-Analyzer was developed that is available from the IEA homepage:

http://www.iea.nl/data.html

The IDB Analyzer is free software with a user friendly interface that creates SPSS syntax based on specifications made by the user (for more details see Sandoval-Hernández & Carrasco, 2020). This syntax can then be applied to an SPSS data file. The IDB Analyzer also offers a merge module for SPSS files that allows merging individual data sets of participating countries (see above).

For analyses in **SAS** and **Stata** the OECD provides macros that can directly be installed and used in the respective analysis program (Keslair, 2020). These macros are available at the OECD homepage.

In order to analyzing PIAAC data using R, the package EdSurvey can be used. This tool takes the replicate design as well as plausible values (if applicable) into account. The installation package is available here: https://www.r-project.org

The use of the R package EdSurvey and its use in analyzing PIAAC data is described by Bailey, Nguyen, Zhang & Lee (2020).

7 Competency levels

A very common analysis for data with competency scores is the presentation of results of a population or subgroups by competency levels. The benchmarks for level classifications are defined in Table 3.

Table 3: Benchmarks for competency levels in PIAAC for Literacy and Numeracy

Level	Benchmark
Below I	≤ 175
I	176 < 226
II	226 < 276
III	276 < 326
IV	326 < 376
V	≥ 376

For Problem solving in technology-rich environments only three competency levels were derived (see Table 4 for competency levels and benchmarks).

Table 4: Benchmarks for competency levels in PIAAC for Problem solving in technology-rich environments

Level	Benchmark
Below I	≤ 240
I	241 < 291
II	291 < 341
III	≥ 341

8 Structure of the data set

Structure of the background information

The background questionnaire used in PIAAC consists of various sections (A to J). The first letter in the variable name indicates the section to which this variable belongs. The number and subsequent letter in the variable name indicates the sequence within one section. For example, variable A_Q01a refers to the first question in section A of the background questionnaire. Respondents did not answer all questions of the questionnaire due to routing processes. For example, respondents currently working were not asked questions about their recent job (see Figure 1).

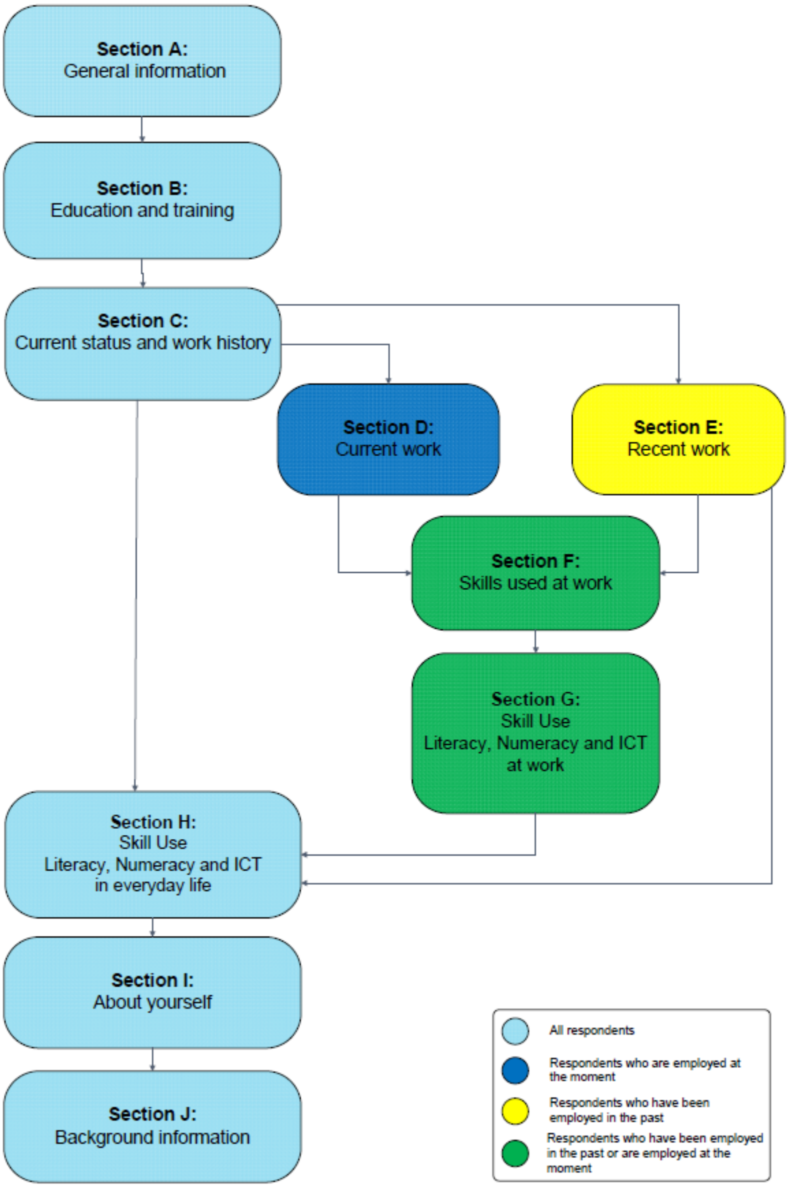
Section A: General information
Section B: Education and training
Section C: Current status and work history
Section D: Current work
Section E: Recent work
Section F: Skills used at work
Section G: Skill Use Literacy, Numeracy and ICT at work
Section H: Skill Use Literacy, Numeracy and ICT in everyday life
Section I: About yourself
Section J: Background information
All respondents
Respondents who are employed at the moment
Respondents who have been employed in the past
Respondents who have been employed in the past or are employed at the moment

Figure 1: Sections and routing in the PIAAC background questionnaire

Source: Zabal et al. (2014)

The international background questionnaire is available at the OECD homepage. The German background questionnaire is available for download at:

https://search.gesis.org/research_data/ZA5845

Additionally to the variables from the background questionnaire, further variables were derived that combine information from one or more background questions. These variables were either derived during the interview process (e.g., C_D05) or afterwards by the PIAAC Consortium during data processing (e.g., NFE12). Both, variables from the background questionnaire and derived variables are specified as "background questionnaire" in the codebook.

While some derived variables are helpful for analyzing the data, it should be noted that for some derived variables, namely B_D01a3DE1, B_D02b3DE1, B_D03b3DE1, and B_D05a3DE1, it is not recommended to use these variables for data analyses. They were derived during the interview before data checks were conducted. The variables are provided in order to better understand the routing of the background questionnaire. If you are interested in using this information, we recommend deriving your own variables based on variables available in the data set.

National adaptation and extensions

The background questionnaire was designed to be comparable across all PIAAC countries. However, some national adaptation had to be made to account for national characteristics such as the different educational systems. Variables capturing the German national adaptations can be identified by the additional letters "DE" in the variable name.

Furthermore, each country had the opportunity to ask some additional questions. In Germany, for example, a question was included where the participant lived before the re-unification of the country. Such national extensions can be identified by the additional letters "DEX" in the variable name.

These national variables are not included in the Public Use File. However, it should be noted that some of these variables were used to derive international variables, such as education variables. The German PIAAC Scientific Use File contains these national variables.

Assessment information

Several variables associated with the cognitive assessment are included in the data set. Scored responses as well as timing variables are specified by the name of the respective competency domain in the codebook. For example, responses and timing information for the literacy domain in the computer branch are referred to as "Literacy (computer)". Plausible values (see above) and PV status are labelled "Scale Scores". Also the total scores for the reading components section are referred to as "Scale Scores".

Sampling and weighting information

Information on sampling and weighting are specified as "Sampling / Weighting" in the codebook. These variables include the final weight, replicate weights and the method used for the replicate approach (VEMETHOD and VEMETHODN). These variables are needed to compute correct standard errors when analyzing the PIAAC data.

IALS trend variables

The International Adult Literacy Survey (IALS) is a predecessor of PIAAC. In order to do trend analyses using IALS and PIAAC data, certain PIAAC variables were recoded to match the categorization of the equivalent variables in IALS. These recoded PIAAC variables are indicated by the variable name with the extension "_T". In the codebook these variables are referred to as "Background questionnaire (trend)".

Some PIAAC countries also participated in the Adult Literacy and Life Skills Survey (ALL). The trend variables in PIAAC are coded so that they can be analyzed with both IALS and ALL data. Germany did not participate in ALL. Thus, trend analyses can only be performed with IALS data. For country-specific problems with the IALS data, that also affect the comparison of the German IALS data with PIAAC, see Gesthuizen, Solga and Künster (2011).

Further information

Further variables are included in the data set, such as information on the interview workflow and about the course of the interview as observed by the interviewer. This information is referred to as "Workflow / logistics" and "Observation module (ZZ questions)".

9 Coarsened variables

The German Scientific Use File contains more detailed information than the German Public Use File accessible on the OECD homepage. However, to ensure data protection, variables in the Scientific Use File containing information on country of origin, disposition codes, number of people in household and of children, and regional information were coarsened.

Country of origin

Information on the origin of the respondents is provided by the variables country of birth (CNT_BIRTH), their parents' country of birth (CNT _BIRTH_M, CNT _BIRTH_F), citizenship and second citizenship (CNT_CITSHIP, CNT_CITSHIP2) and the country in which the highest qualification was obtained (CNT_H). These variables are coarsened. Countries that represent groups of foreign residents with less than 50 000 inhabitants in Germany are aggregated with similar countries according to the categorization of countries used in the German Microcensus (see Table 5). Countries were coded in ISO 3166. The category of a group of aggregated countries received the ISO-Code of one of the countries included in this category. For example the aggregated category that combines Denmark, Finland, and Sweden received the code 208, which is the country code for Denmark in the ISO 3166 categorization.

Please note that when combining the German PIAAC Scientific Use File with the Public Use File, country information in the labels might be lost. Country codes used for aggregated categories in the German PIAAC Scientific Use File may represent a different set of countries than the same code in the Public Use File. We therefore recommend using the syntax provided in this user guide (see "Merging with the PIAAC Public Use File" above) when merging the German Scientific Use File with the Public Use File.

Table 5: Country codes and aggregated categories for country information

008	Albania; Andorra; Moldavia; Monaco; San Marino; Vatican City State; Belarus
208	Denmark; Finland; Sweden
705	Estonia; Slovenia; Latvia; Lithuania; Malta; Cyprus
528	Belgium; Luxembourg; Netherlands
756	Iceland; Liechtenstein; Norway; Switzerland
203	Slovakia; Czech Republic
826	Ireland; United Kingdom
688	Montenegro; Serbia
226	Equatorial Guinea; Gabon; Cameroon; Democratic Republic of the Congo; Congo; Sao Tome and Principe; Sudan; Chad; Central African Republic; Ethiopia; Burundi; Djibouti; Eritrea; Kenya; Comoros; Madagascar; Mauritius; Rwanda; Seychelles; Somalia; United Republic of Tanzania; Uganda; Angola; Botswana; Lesotho; Malawi; Mozambique; Namibia; Zambia; Zimbabwe; South Africa; Swaziland

288	Benin; Burkina Faso; Côte d'Ivoire; Gambia; Ghana; Guinea; Guinea-Bissau; Cape Verde; Liberia; Mali; Mauritania; Niger; Nigeria; Senegal; Sierra Leone; Togo
788	Algeria; Libya; Tunisia; Egypt
076	Argentina; Bolivia; Brazil; Chile; Ecuador; Guyana; Colombia; Paraguay; Peru; Suriname; Uruguay; Venezuela (Bolivarian Republic of); all other South American countries
124	Canada; Antigua and Barbuda; Bahamas; Barbados; Belize; Costa Rica; Dominica; Dominican Republic; El Salvador; Grenada; Guatemala; Haiti; Honduras; Jamaica; Cuba; Mexico; Nicaragua; Panama; Saint Kitts and Nevis; Saint Lucia; Saint Vincent and the Grenadines; Trinidad and Tobago; all other Central American countries and the Caribbean
268	Armenia, Azerbaijan, Georgia; Armenia; Azerbaijan; Georgia; Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan; Kyrgyzstan; Tajikistan; Turkmenistan; Uzbekistan; Mongolia
356	Sri Lanka; India
050	Bangladesh; Bhutan; Brunei Darussalam; Indonesia; Cambodia; Lao People's Democratic Republic; Malaysia; Maldives; Myanmar; Nepal; Pakistan; Philippines; Singapore; Timor-Leste
410	Japan, Taiwan; Democratic People's Republic of Korea; Republic of Korea
760	Yemen; Bahrain; Israel; Jordan; Qatar; Oman; United Arabian Emirates; Saudi-Arabia; Syria; all other countries of the Middle East
036	Australia; Fiji; Micronesia, Federated States of; New Zealand; Niue; Papua New Guinea; Solomon Islands; Cook Islands; Kiribati; Marshall Islands; Nauru; Palau; Samoa; Tonga; Tuvalu; Vanuatu; all other countries

Disposition codes

Disposition codes indicate which part(s) of the interview a respondent has completed and, if applicable, reasons for not completing the full interview. Reasons for not completing the interview can be related to the respondents' health or migration status. In order to protect this sensitive information disposition codes included in the German Scientific Use File were coarsened. Included are the two disposition codes DISP_MAIN_C and DISP_MAINWRC_C that combine various reasons of language problems and disabilities in one category each.

Number of people in household and number of children

The number of people in households and the number of children can be used to identify respondents. Therefore, this information was top-coded. The dataset includes the information on the number of people in the household, top-coded at 6 and at 10 (J_Q01_C and J_Q01_C10, J_Q01_T1, J_Q01_T2) as well as information on the number of children, top-coded at 4 and at 6 (J_Q03b_C and J_Q03b_C6).

Regional information

Municipality size was coarsened and is published in eight categories. The first five categories represent categories for municipality size used by the German Microcensus. For some federal states, municipality size had to be coarsened in order to prevent the identification of single municipalities. The specific federal states are indicated in brackets in category 6 and 7. Category 8 is reserved for the federal state of Bremen (incl. Bremerhaven). For these federal states information in the corresponding categories 1 to 5 are suppressed (see Table 6).

Table 6: GKPOL – Political municipality size in 8 categories

1	1 to 4 999 inhabitants
2	5 000 to 19 999 inhabitants
3	20 000 to 99 999 inhabitants
4	100 000 to 499 999 inhabitants
5	500 000 to 99 999 999 inhabitants
6	1 to 19 999 inhabitants (North Rhine-Westphalia, Hesse, Saxony-Anhalt)
7	20 000 to 499 999 inhabitants (Saarland, Brandenburg, Mecklenburg-Western Pomerania, Saxony, Saxony-Anhalt, Thuringia)
8	Bremen (unique key)

Regional information is also included in certain variables used for weighting processes, such as ID_PSU and STRAT_PSU. In order to ensure that no respondent can be identified, the values of the original variables were randomly reassigned to each person.

10 Characteristics of selected variables

German Federal States

The international variable REG_TL2 provides information on the German Federal States as a string variable. An additional numeric variable (Federal States) is offered to allow easier data handling when working with the German Scientific Use File alone.

Hypothetical years of education (YRSQUAL)

The variable YRSQUAL was created based on the average or most usual time that it takes to complete a qualification as indicated by different national sources such as the Federal Statistical Office and national education experts.

Income

Respondents were asked to report their gross income in their current job (D_Q16a to D_Q18c2). They could choose in which interval they wanted to declare their income (e.g., per hour, per day, per week). In case respondents did not want to provide their exact income, they were able to answer this question in broad categories.

Due to these different options for providing income information, the variables D_Q16a to D_Q18c2 contain numerous missing data. For easier data handling, several derived variables where created by the PIAAC Consortium and the OECD that combine this information (see Table 7).

Table 7: Derived income variables

Variable name	Variable label
EARNHR	Hourly earnings excluding bonuses for wage and salary earners (derived)
EARNHRDCL	Hourly earnings excluding bonuses for wage and salary earners, in deciles (derived)
EARNHRPPP	Hourly earnings excluding bonuses for wage and salary earners, PPP corrected $US (derived)
EARNHRBONUS	Hourly earnings including bonuses for wage and salary earners (derived)
EARNHRBONUSDCL	Hourly earnings including bonuses for wage and salary earners, in deciles (derived)
EARNHRBONUSPPP	Hourly earnings including bonuses for wage and salary earners, PPP corrected $US (derived)
EARNMTH	Monthly earnings excluding bonuses for wage and salary earners (derived)
EARNMTHPPP	Monthly earnings excluding bonuses for wage and salary earners, PPP corrected $US (derived)
EARNMTHSELFPPP	Monthly earnings for self-employed, PPP corrected $US (derived)

EARNMTHBONUS	Monthly earnings including bonuses for wage and salary earners (derived)
EARNMTHALL	Monthly earnings including bonuses for wage and salary earners and self-employed (derived)
EARNMTHALLDCL	Monthly earnings including bonuses for wage and salary earners and self-employed, in deciles (derived)
EARNMTHALLPPP	Monthly earnings including bonuses for wage and salary earners and self-employed, PPP corrected $US (derived)
EARNMTHBONUSPPP	Monthly earnings including bonuses for wage and salary earners, PPP corrected $US (derived)

Source: Codebook of the PIAAC Public Use File (PUF).

Industry (ISIC4)

Similarly to the respondents' occupation, the industry sector of the current or the recent job was collected using open questions. This information was re-coded by the DPC according to the International Standard Industrial Classification of All Economic Activities, Revision 4 (ISIC4, United Nations Statistics Division, 2013) into four digits ISIC. These string variables can include letters (1-digit codes) or numbers (2-4 digit codes).

Job Requirements Approach (JRA, Skill use)

The job requirements approach, a set of self-reported questions, was developed for the PIAAC background questionnaire based on previous work by Felstead, Gallie, Green, and Zhou (2007). Respondents had to answer questions regarding cognitive and non-cognitive skills used at work and outside of work. From this information a number of skill use scales were constructed by the PIAAC Consortium and the OECD using item response theory, more specifically the generalized partial credit model (GPCM) and Warm's mean weighted likelihood estimation (WLE). Respondents that answered all questions regarding one skill use domain with "never" (all-zero-response) were excluded from these indices. For detailed information see OECD (2013b).

A total of 13 skill use indices were derived (see Table 8). For each skill use domain the data set includes three variables: the mean score (e.g., READWORK), its standard error (e.g., READWORK_SE) and skill use indices in categories 0-5 (e.g., READWORK_WLE_CA). The mean score and standard errors are standardized with a mean equal to 2 and a standard error equal to 1 across the OECD countries participating in PIAAC.

Additional skill use indices were derived directly from individual items in the questionnaire. These are problem solving, co-operative skills, self-organizing skills, physical skills and dexterity.

Table 8: Composition of JRA scales

Variable name	Index name	Variables used to derive index
TASKDISC	Index of use of task discretion at work	D_Q11a, D_Q11b, D_Q11c, D_Q11d
LEARNATWORK	Index of learning at work	D_Q13a, D_Q13b, D_Q13c
INFLUENCE	Index of use of influencing skills at work	F_Q02b, F_Q02c, , F_Q02e, F_Q03b, F_Q04a, F_Q04b, (excluded F_Q02d)
PLANNING	Index of use of planning skills at work	F_Q03a, F_Q03b, F_Q03c
READWORK	Index of use of reading skills at work (prose and document type texts)	G_Q01a, G_Q01b, G_Q01c, G_Q01d, G_Q01e, G_Q01f, G_Q01g, G_Q01h
WRITWORK	Index of use of writing skills at work	G_Q02a, G_Q02b, G_Q02c, G_Q02d
NUMWORK	Index of use of numeracy skills at work (basic and advanced)	G_Q03b, G_Q03c, G_Q03d, G_Q03f, G_Q03g, G_Q03h
ICTWORK	Index of use of ICT skills at work	G_Q05a, G_Q05c, G_Q05d, G_Q05e, G_Q05f, (excluded G_Q05g), G_Q05h
READHOME	Index of use of reading skills at home (prose and document type texts)	H_Q01a, H_Q01b, H_Q01c, H_Q01d, H_Q01e, H_Q01f, H_Q01g, H_Q01h
WRITHOME	Index of use of writing skills at home	H_Q02a, H_Q02b, H_Q02c, H_Q02d
NUMHOME	Index of use of numeracy skills at home (basic and advanced)	H_Q03b, H_Q03c, H_Q03d, H_Q03f, H_Q03g, H_Q03h
ICTHOME	Index of use of ICT skills at home	H_Q05a, H_Q05c, H_Q05d, H_Q05e, H_Q05f, H_Q05g, H_Q05h
READYTOLEARN	Index of readiness to learn	I_Q04b, I_Q04d, I_Q04h, I_Q04j, I_Q04l, I_Q04m

Source: OECD (2013a).

Migration background (IMGEN)

The derived variable IMGEN provides information on natives as well as first and second generation immigrants based on information on country of birth of the respondents' parents. An additional category (category 4) includes those respondents with just one parent born in a foreign country. Thus, these respondents are neither native nor a first or second generation immigrant. Since no clear categorization as native or first or second immigrant is possible for this group of people, this category is defined as missing values. However, depending on research focus, this category may be of interest for the user. The following SPSS syntax can be used to define IMGEN category 4 as a valid category:

```
missing values IMGEN ().

missing values IMGEN (9).
```

Occupation (ISCO08)

The respondents' current or recent occupation and their parents' occupation when the respondent was 16 years old were recorded using a set of open questions each. This information was re-coded by the Data Processing and Research Center (DPC) according to the International Standard Classification of Occupations 2008 (ISCO08, International Labour Organization, 2012) into four digits ISCO. Information used for coding was the exact title of the occupation (D_Q01a), a description of the occupation, and the most important tasks (D_Q01bDE1).

Please note that the ISCO codes are string variables and contain leading zeros, i.e. the codes can start with values from 0-9.

Recoded variables (REC)

Certain variables were recoded to include additional information that respondents provided in open-end questions. For example, when a respondents chose category "Other qualification" when asked about their highest qualification obtained (e.g., B_Q01aDE1), they had the opportunity to describe their highest qualification obtained in an open-end question (B_S01a1DE1). When possible, the information provided in B_S01a1DE1 was recoded into the categories used in B_Q01aDE1 and then added to this variable by recoding it into a new variable B_Q01aDE1_REC. Thus, this variable contains more information than the original variable. These recoded variables were therefore included in the data set instead of their originals. These recoded variables were also used to derive further variables, such as B_Q01a, EDCAT6, EDCAT7, and EDCAT8.

Routing for national adaptations and extensions

The German Scientific Use File provides national variables identified by the letters "DE" or "DEX" (see "National adaptation and extensions" above). The routing for these variables may differ slightly from the routing of the international variables. For example, in the German background questionnaire all respondents were asked whether the main reason for choosing the currently pursued qualification was job related. In the international background questionnaire, only respondents 20 years and older had to provide this information. All international variables contain values following the international routing (to align with the Public Use File), although the national routing may suggest otherwise.

Time of day and weekday of the interview

During the interview date and time were automatically recorded. From this information the time of day (Time_of_day) and the weekday (Weekday) were extracted and are available in the German Scientific Use File. This information allows analyses, for example, on performance and time of interview.

Time_of_day and Weekday are based on the information about the time when the case was initialized and the interview process started. The assessment followed the background questionnaire and started, on average, 45 minutes after case initialization.

11 References

Bailey, P., Nguyen, T., Zhang, T. & Lee, M. (2020). Using EdSurvey to Analyze PIAAC Data. In D. B. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment: Analyzing PIAAC data*. Serie: Methodology of Educational Measurement and Assessment. New York: Springer. doi: 10.1007/978-3-030-47515-4

Felstead, A., Gallie, D., Green, F., & Zhou, Y. (2007). *Skills at work, 1986 to 2006*. Cardiff: ESRC Research Centre on Skills, Knowledge and Organizational Performance.

Gesthuizen, M., Solga, H., & Künster, R. (2011). Context matters: Economic marginalization of low-educated workers in cross-national perspective. *European Sociological Review*, 27(2), 264-280. doi: 10.1093/esr/jcq006

International Labour Organization. (2012). *International standard classification of occupations ISCO-08*. Genf: International Labour Organization.

Keslair, F. (2020). Analysing PIAAC Data with Stata. In D. B. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment: Analyzing PIAAC data*. Serie: Methodology of Educational Measurement and Assessment. New York: Springer. doi: 10.1007/978-3-030-47515-4

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177-196.

OECD. (2019). Technical report of the Survey of Adult Skills. 3. Edition. Paris: OECD. Available at http://www.oecd.org/skills/piaac/publications/PIAAC_Technical_Report_2019.pdf

OECD. (2009). *PISA data analysis manual: SPSS, second edition (OECD Ed.)*. Paris: OECD Publishing.

OECD. (2013a). *The Survey of Adult Skills: Reader's companion*. Paris: OECD Publishing.

OECD. (2013b). *Technical report of the Survey of Adult Skills*. Paris: OECD Publishing.

Rammstedt, B., Ackermann, S., Helmschrott, S., Klaukien, A., Maehler, D.B., Martin, S., Massing, N., & Zabal, A. (Eds.) (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich: Ergebnisse von PIAAC 2012*. Münster: Waxmann.

Rammstedt, B., Martin, S., Zabal, A., Konradt, I., Maehler, D., Perry, A., . . . Helmschrott, S. (2016). *Programme for the International Assessment of Adult Competencies (PIAAC), Germany - Reduced Version* (Version 2.2.0) [ZA5845]. GESIS Data Archive, Cologne. doi: 10.4232/1.12660

Sandoval -Hernández, A. & Carrasco, D. (2020). Analysing PIAAC data with the IDB Analyzer (SPSS and SAS). In D. B. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment: Analyzing PIAAC data*. Serie: Methodology of Educational Measurement and Assessment. New York: Springer. doi: 10.1007/978-3-030-47515-4

Statistics Canada. (2002). *The Adult Literacy and Life Skills Survey, 2003. Public use microdata file. User's manual.* Ottawa: Statistics Canada.

United Nations Statistics Division. (2013). *Detailed structure and explanatory notes ISIC Rev. 4*. Retrieved 12.07.2013, from http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27

Von Davier, M., Gonzalez, E. J., & Mislevy, R. J. (2009). *What are plausible values and why are they useful?* IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2, 9-36.

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). *PIAAC Germany 2012: Technical report.* Münster: Waxmann. Available at https://www.gesis.org/fileadmin/piaac/Downloadbereich/TechnicalReport-ebook.pdf