What You Need to Consider Before Working with PIAAC Data

AIR PIAAC Team

Sampling Weights

Sampling weights are designed to make the data representative of the target population by compensating for the disproportionate sampling of subgroups and non-coverage, reducing sampling errors by making use of known data for the population, minimizing biases arising from differences between respondents and non-respondents, and facilitating the estimation of variances through the use of the replication approach.

- **Final Weight:** A final weight was required for all sampled persons with a completed background questionnaire and those who could not complete the background questionnaire for literacy-related reasons, but for whom age and gender were collected. The final weight is a result of various adjustment procedures. To calculate estimates representative of the U.S. PIAAC population, you need to select the final weight, which is called **SPFWT0**.
- **Replicate Weights:** Participating countries have used one of four different replication schemes. The U.S. used the paired jackknife, or JK2, method with 80 replicate weights. To calculate representative standard errors for the U.S. PIAAC population, you need to select the replicate weights that are associated with the final weight. The replicate weights associated with the final weight SPFWT0 are **SPFWT1 through SPFWT80**.

Plausible Values

Plausible values (PVs) are a statistical means to replicate a probable score distribution that summarizes how well each respondent answered a small subset of the assessment items; and, how well other respondents from a similar background performed on the rest of the assessment item pool. Each individual case in the PIAAC dataset has a set of ten PVs for each proficiency domain (literacy, numeracy, problem solving in technology-rich environments), and all ten PVs must be used together to estimate proficiency. On the data file, PVs for literacy are labeled **PVLIT1 through PVLIT10**, PVs for numeracy are labeled **PVNUM1 through PVNUM10**, and PVs for problem solving in technology-rich environments are labeled **PVPSL1 through PVPSL10**.

For accurate estimations involving proficiency scores, calculations must account for both the sampling error component, and the variance due to imputation of the proficiency scores. To account for the sampling error component, you must use the final weight and the corresponding 80 replicate weights. To account for the imputation variance, you must use all ten plausible values.

Missing Data

Missing data can occur when some of the adults selected in the sample are not accessible or refuse to participate, when they fail to respond to a particular survey item, or, because data collected from the sampled adults are contaminated or lost during or after the data collection phase. All missing data for the PIAAC Background Questionnaire are marked in the dataset as valid skips, don't know, refused, or not stated/inferred. No Background Questionnaire data in the U.S. national public-use data file or restricted-use data file were imputed. All missing assessment item responses are marked as missing, and no answers were imputed. A small proportion of the sample did not respond to the PIAAC background questionnaire as a result of language difficulties or learning or mental disabilities. These cases do not have plausible values for any of the domains. In addition, respondents to paper-based assessment were routed out of the problem solving in technology-rich environments assessment. These cases do not have plausible values for the problem solving in technology-rich environments domain.

Available Data Files

- **OECD International U.S. public-use file (PUF)** in SPSS and SAS formats is available for download on the OECD PIAAC website along with the PUF for every country participating in PIAAC, except Australia and Cyprus. Directions on how to obtain the Australian data can be found on the OECD website, while the Cyprus file is available for download from GESIS Data Catalog. Some detailed data on respondents is suppressed or categorized in all of the public use data sets.
  Available from OECD:
  http://www.oecd.org/site/piaac/publicdataandanalysis.htm
- **NCES U.S. National PUF** in ASCII, SPSS, and SAS formats is available for download on the NCES PIAAC website. This file includes U.S.-only variables, plus all variables in the international PUF. Some detailed data on respondents is suppressed or categorized in this data set.
  Available from NCES:
  http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014045
- **NCES U.S. National restricted-use file (RUF)** in ASCII, SPSS, and SAS formats is not available for download on the NCES PIAAC website. To access the restricted-use data, the restricted-use license has to

be applied for and obtained from NCES. More information on the process is available at: http://nces.ed.gov/pubsearch/licenses.asp. The file includes U.S.-only variables, in additional to all variables in the international and national PUFs. This file includes all of the variables that were suppressed or categorized in the PUFs in their original level of detail.

- **OECD U.S. National synthetic restricted-use file (S-RUF)** in SAS and SPSS formats is available to researchers outside of the U.S. to develop and test computer code for the analysis of PIAAC data on the NCES U.S. National restricted-use file. As the data on the S-RUF is synthetic, no conclusions can be drawn from the output generated. Researchers wishing to run analysis on the actual U.S. RUF are able to submit their SAS, SPSS, or Stata code to piaac@air.org, where the requested analyses will be run using on the real U.S. RUF. After undergoing a confidentiality review, the approved output will be returned to the researcher.
  Available from OECD: http://www.oecd.org/site/piaac/publicdataandanalysis.htm

Demographic Variables

Below are some common demographic categories and the common variables that are used in analyses:
Gender: GENDER_R;
Age: AGEG10LFS for age in 10-year intervals or AGEG5LFS for age in 5-year intervals; National RUF only: AGE_R for continuous age;
Race: RACETHN_4/5CAT for 4- or 5- category race variable; National RUF only: RACETHN_6CAT for 6-category race variable;
Income: EARNMTHALLDCL for monthly earnings in deciles; National PUF: EARNMTHALLPPPUS_C: monthly earnings, purchasing power parity (PPP) corrected, topcoded; National RUF only: EARNMTHALL/EARNMTHALLPPP: monthly earnings/PPP corrected in $US. PPP corrected income variable is used for cross-country comparisons;
Educational Attainment: EDLEVEL3 for three categories of educational attainment; EDCAT6/7/8 for 6, 7, or 8 categories of educational attainment; National PUF: B_Q01A_C or B_Q01AUS_C for three categories of educational attainment;
Nativity/Immigration: J_Q04A for whether the respondent was born/not born in country;
Employment Status: C_D05 for 3-category employment and labor force participation status.

Data Tools

- **International Data Explorer (IDE):** a user-friendly, online tool that can be used for basic analyses and takes into account the plausible values and sampling weights for you (U.S. IDE: http://nces.ed.gov/surveys/international/ide/; OECD IDE: http://piaacdataexplorer.oecd.org/ide/idepiaac/)
- **IEA International Database (IDB) Analyzer:** a Windows program used in conjunction with SPSS that provides a point-and-click interface to merge the micro-data files of the participating countries and create SPSS syntax that takes into account the plausible values and sampling weights (http://www.iea.nl/data.html)
- **SAS macro (PIAAC_TOOL) and data analysis manual:** macro used to do analysis taking into account the plausible values and sampling weights (http://www.oecd.org/site/piaac/publicdataandanalysis.htm)
- **Stata macro (PIAACTOOLS) and data analysis manual:** macro used to do analysis taking into account the plausible values and sampling weights (http://www.oecd.org/site/piaac/publicdataandanalysis.htm)

Further Considerations for Data Analysis

- The data was collected from samples of people who are representative of groups, not individuals
- The data is cross-sectional and only captures one point in time. However, several measures, including those for literacy and numeracy, can be compared with rescaled data from the 1994 IALS and 2003 ALL studies.
- The data was derived from non-experimental research, so data should only be analyzed in terms of non-causal relationships.
- To produce meaningful, reliable results, a sample size of 62 cases per analytic group or subgroup is recommended.

Additional Resources on PIAAC Data and Analysis

- Distance Learning Dataset Training (DLDT) Common Modules: Online training modules allowing you to learn more the statistical procedures and methods for proper analysis of NCES datasets.
- Distance Learning Dataset Training (DLDT) PIAAC Modules: Online training modules that will allow you to learn more about PIAAC at your own pace.
- OECD *Technical Report of the Survey of Adult Skills (PIAAC).*
- OECD *The Survey of Adult Skills: Reader's Companion*.

Specific questions others have had about PIAAC data and analysis, and responses to these questions can be found on the Q&A page of the PIAAC Gateway.