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.

Demographic Variables

Below are some common demographic categories and the common variables that are used in analyses. The list applies to both household and prison files unless noted:
Gender: GENDER_R;
Age: International PUF: AGEG10LFS for age in 10-year intervals or AGEG5LFS for age in 5-year intervals; National PUF: AGEG10LFSEXT or AGEG5LFSEXT for age bands extended to include ages over 65; 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 (Household Only): 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: 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 (Household Only): C_D05 for 3-category employment and labor force participation status.

Sample Sizes

- For analyses involving plausible values:
  - Unweighted cell size of 62 cases or more is enough to report statistics including means, standard errors, standard deviations, a set of percentiles, a set of proficiency level percentages, and do regression analysis. This 62-case limit is used in NCES reporting and data tools, such as the International Data Explorer (IDE).
  - If an unweighted cell size for analyses is between 30 and 61 cases, it is recommended that the estimate is marked with a "!" sign and added note, such as "! Interpret data with caution. The sample size for this estimate is between 30 and 61 cases." This 30-case limit is based on the way OECD report their analysis.
  - If an unweighted cell size for analysis is less than 30, it is recommended suppressing the results, using the "‡" sign instead of the estimate, and using a note, such as "‡ Reporting standards not met."
- For analyses not involving plausible values:
  - Unweighted cell size of 30 cases or more is enough to report statistics including means, standard errors, standard deviations, a set of percentiles, a set of percentages, and do regression analysis. For percentages, the results are reportable if the numerator is above 2 and the denominator is 30 or above.
  - If the cells sizes of the analysis groups do not meet these standards, suppressing the data using a note such as "‡ Reporting standards not met." is recommended.
- Note that cell size for an analysis is the unweighted N size for each reporting subgroup. For example, if the reporting is for average score results of male or females by nativity, the unweighted sample size to check would be for native-born males, native-born females, non-native-born males, and non-native-born females.
- In regression analysis, the cell size is the unweighted N size of each reporting subgroup for each variable in the regression. For example, for a regression of gender and nativity status on literacy, the unweighted sample size to check would be for males, females, native-born and non-native born.
- After the estimates pass the sample size check an additional check is to look at the coefficient of variation (CV), i.e. the standard error divided by the estimate expressed as a percentage. For estimates with CVs between 30% and 50%, NCES standard is to mark it with "!" sign and adding a note "! Interpret data with caution. The coefficient of variation (CV) for this estimate is between 30 and 50 percent." For estimates with CVs greater than 50%, NCES standards is to suppress the data, use the "‡" sign instead of the estimate and use a note such as "‡ Reporting standards not met."

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. 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.
- The U.S. data from 2014 is not nationally representative and cannot be analyzed on its own; the combined 2012/14 data should be used for analysis.
- The U.S. 2014 prison data is representative of the incarcerated adults in the United States and can be compared with the U.S. 2012/14 data on the household population along common variables and across skills measures. It is not internationally comparable.
- The U.S. household sample ages 16-65 can be compared internationally. U.S. 2012/14 data includes adults 66-74 that need to be excluded when international comparisons are conducted.

Suggested Citation: AIR PIAAC Team (2019, January). What You Need to Consider Before Working with PIAAC Data. Retrieved [date] from http://piaacgateway.com/s/WorkingWithPIAACData_updated-092816.pdf