

Chapter 14: Sampling Design

Leyla Mohadjer, Tom Krenzke and Wendy Van de Kerckhove, Westat

This chapter presents information about the PIAAC Main Study sample design and selection results. Participating countries were required to develop their sample design and selection plans according to the standards provided in the PIAAC Technical Standards and Guidelines (TSG) and to submit their plans to the Consortium for approval. The sample design plans included information about sampling frames and their coverage, providing descriptions of the national sample designs that included stages of sampling, probabilities of selection, sampling units and sample sizes. The sample selection plans included detailed information about the processes for sample selection at each stage of sampling. In addition, the countries were required to complete and submit quality control sample selection forms to the Consortium to verify that the sample selection was conducted in an unbiased and randomized way consistent with PIAAC standards.

The target population for PIAAC consists of all noninstitutionalized adults between age 16 and 65 (inclusive) who reside in the country (meaning their usual place of residency is in the country) at the time of data collection. Countries were allowed to expand the target population to include additional subpopulations of interest to the country as long as they followed the TSG on such supplementation. Section 14.1 provides more detail on the PIAAC target population and the national target populations if expanded beyond the PIAAC standard definition. Section 14.2 contains information about the sources of country sampling frames and their coverage of the target population.

The TSG allowed each country to choose a sample design and selection approach that is most optimal and cost effective as long as the design applies full selection probability methods to select a representative sample from the PIAAC target population. Descriptions of the standard PIAAC and national sample designs and probabilities of selection are given in section 14.3. The definition of sampling units and sample selection methods are provided in section 14.4. Section 14.5 contains the PIAAC target sample sizes and describes the process applied to determine the initial sample sizes. Sample selection results and a summary of the sampling quality control procedures are given in section 14.6 and section 14.7, respectively. Finally, section 14.8 provides a brief description of the incentive plans for PIAAC.

14.1 Target population and sampling frame

A clear and precise definition of the target population is necessary to ensure that the population of interest is adequately covered by each participating country and to maintain consistency and comparability across countries. The PIAAC target population consists of all noninstitutionalized adults between age 16 and 65 (inclusive) who reside in the country (usual place of residency is in

the country) at the time of data collection. Adults were to be included regardless of citizenship, nationality or language (standard 4.1.1). The target population excludes adults in institutional collective dwelling units (or group quarters) such as prisons, hospitals and nursing homes, as well as adults residing in military barracks and military bases. However, full-time and part-time members of the military who do not reside in military barracks or military bases are included in the target population.

Adults in other noninstitutional collective dwelling units (or group quarters), such as workers' quarters or halfway homes, are also included in the target population. This includes adults living at school in student group quarters such as a dormitory, fraternity or sorority. Adults who were unable to complete the assessment because of a hearing impairment, blindness/visual impairment or physical disability are considered in scope; however, they were excluded from PIAAC response rate calculations because the assessment does not accommodate such situations.

The target population does not cover the entire geography area for the following countries:

- Belgium – The target population consists of Flanders, which is in the northern portion of the country.
- Cyprus¹ – The target population consists of the area under the effective control of the Government of the Republic of Cyprus, which includes the districts of Nicosia (part), Limassol, Larnaca (part), Paphos and Famagusta (part).
- Russian Federation² – The target population does not include Moscow or Moscow Region.

Some countries expanded the target population to include additional subpopulations of interest to the country. These country-specific supplemental samples, approved by the Consortium, are presented in Table 14-1 below.

Table 14-1: Country-specific samples

Country	Specific samples
Australia	Persons aged 15 and 66-74
Denmark	PISA 2000 survey respondents aged 26-27

Some countries elected to oversample portions of the target population. The oversamples approved by the Consortium are presented in Table 14-2 below.

¹ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

² Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 14-2: Countries conducting oversampling

Country	Group oversampled
Australia	Persons living in certain states/territories
Canada	Individuals aged between 16 and 24 inclusive in British Columbia; Linguistic minorities (English in Québec, French elsewhere) in New Brunswick, Québec, Ontario and Manitoba; Métis in Ontario; Aboriginal individuals in Québec, Ontario, Manitoba, Saskatchewan, Alberta, British Columbia and Yukon Territory; and Recent immigrants (living in Canada since 2002 or after) in Québec, Ontario, Alberta and British Columbia
Czech Republic	Persons aged 16-29
Denmark	Persons aged 55-65 years and immigrants 16-65
Germany	Persons aged 26-55 living in former East Germany or former East Berlin ¹
Poland	Persons aged 19-26

¹ For national purposes; not included in the international data.

14.2 Sampling frames and their coverage

The sampling frame is the list from which the sample is selected, so the quality of the sampling frame affects the quality of the sample. In addition, adequate information on the frame must be available to conduct sampling, data collection, weighting and nonresponse bias analyses. Most countries with multiple stages of selection had specified multiple frames. Those frames were reviewed by the Consortium to ensure they included sufficiently reliable information for sampling individual units and ultimately locating individuals for the interview and assessment. Section 14.2.1 provides information about the sampling frames used at each stage of selection, while section 14.2.2 contains information about the coverage of these frames.

In PIAAC, the noncoverage rate, combined over all stages of sampling, could not exceed 5% (standard 4.1.2). Thus the sampling frames for each country were required to include 95% or more of the standard PIAAC target population. Frame noncoverage rates (see section 14.2.2) were limited as much as possible so that no extensive biases are introduced as a result of noncoverage of some subgroups of the population.

14.2.1 Sampling frames

PIAAC standards require that sampling frames be up to date and include only one record for each member of the target population. Countries had to examine their sampling frames and eliminate duplicate records when lists were combined to create a sampling frame. Countries were required to assess the extent of duplication and the proportion of out-of-scope units on the frame and, if necessary, develop a plan to correct these problems. In addition, countries also evaluated and developed plans to address any noncoverage in the frame that was not addressed in the documentation of country-specific exclusions (see Table 14-6). The methodology used to create these frames was also reviewed by the Consortium.

Multistage sample designs required a sampling frame for each stage of selection. Some countries used national population registries as sampling frames, which contain useful variables for stratification, weighting and nonresponse bias analyses. If the country had a list of residents that was of sufficient quality, no frame of households or household sampling was necessary.

However, some countries' lists of residents used for the study did not completely cover the PIAAC target population (e.g., the lists may have excluded nonnationals/noncitizens), complicating their use as a sampling frame. See Table 14-3 for the full list of sampling frames employed by countries with population registry samples.

Table 14-3: Sampling frames for countries with population registry samples

Country	Sampling frame		
	Stage 1	Stage 2	Stage 3
Austria	Population registry, 2011		
Denmark	Population registry, 2011		
Estonia	Population registry, 2011		
Finland	Statistics Finland's population database (based on the Central Population Register), 2011		
Flanders (Belgium)	Population registry, 2011		
Germany	German Census Bureau frame of communities, 2011	Local population registries, 2011	
Italy	National Statistical Institute of Italy frame of municipalities, 2011	Household registries held by municipalities, 2011	Population registries, 2011; combined with field enumeration
Japan	Resident registry, 2011	Resident registry, 2011	
Netherlands	Population registry, 2011		
Norway	Population registry, 2011		
Poland	Population registry, 2011	Population registry, 2011	
Slovak Republic	Population registry, 2011	Population registry, 2011	
Spain	Population registry, 2011	Population registry, 2011	
Sweden	Population registry, 2011		

■ indicates there is no such stage in the country's sample design.

Some countries have access to master samples used for national surveys. For example, Australia has a master sample of dwelling units (DUs) already in use by governmental surveys that was also used for PIAAC. Similarly, Australia and France have master samples of area primary sampling units (PSUs). See Table 14-4 for more information on how master samples were employed by participating countries.

Table 14-4: Sampling frames for countries using master samples

Country	Sampling frame			
	Stage 1	Stage 2	Stage 3	Stage 4
Australia	Bureau of Statistics population survey master sample, 2006	Bureau of Statistics population survey master sample, 2006	Bureau of Statistics population survey master sample, 2006	Field enumeration
France	Master sample from census data file, 1999	Individual taxation file, 2011		

For multistage area sample designs in which a registry is not being used, listing procedures are necessary to create a frame of households within the selected geographic clusters. A frame of geographic clusters can be formed by combining adjacent geographic areas, respecting their population sizes and taking into consideration travel distances for interviewers. Table 14-5 contains sampling frames for the remaining countries without registries using area sample designs for PIAAC. The exception is that Cyprus³ is included in Table 14-5 among the countries without population registries, even though it did not use an area sample design, Cyprus did not require listing procedures because its sample frame for the first stage was a list of households from the Statistical Service Census 2001, updated with information from the 2010 Electricity Authority Household Registry.

³ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 14-5: Sampling frames for countries without population registries and master samples

Country	Sampling frame			
	Stage 1	Stage 2	Stage 3	Stage 4
Canada	Short-form Census returns and National Household Survey returns for some oversamples, 2011	Short-form Census returns and National Household Survey returns for some oversamples, 2011	Field enumeration	
Cyprus ⁴	List of households from the Statistical Service Census 2001, updated with information from the 2010 Electricity Authority Household Registry	Field enumeration		
Czech Republic	Territorial Identification Register of Buildings and addresses (UIR-ADR), 2010	Territorial Identification Register of Buildings and addresses (UIR-ADR), 2010	Field enumeration	Field enumeration
England (UK)	Royal Mail list of UK Postal Sectors, 2011	Royal Mail PAF residential file, 2011	Field enumeration	Field enumeration
Ireland	Small Area classifications, 2006	2011 Census	Field enumeration	
Korea	2010 Census	2010 Census	Field enumeration	
Northern Ireland (UK)	NI(POINTER) database, 2011	Field enumeration	Field enumeration	
Russian Federation ⁵	Federal State Statistics Service, data of the national survey organizations, 2010	Federal State Statistics Service, data of the national survey organizations, 2010	Official data of urban districts, 2010	Field enumeration
United States	Census Bureau Population Estimates, 2008	2000 Census Bureau Summary File 1 (SF1), 2000; updated with data from the United States Postal Service 2010	Field enumeration	Field enumeration

■ indicates there is no such stage in the country's sample design.

14.2.2 Noncoverage of the target population

As mentioned earlier, the noncoverage rate for PIAAC, combined over all stages of sampling, may not exceed 5% (standard 4.1.2), and thus the sampling frames for each country were required to include 95% or more of the standard PIAAC target population. All exclusions to the core PIAAC target population, whether or not they exceed the threshold, were reviewed by the Consortium. Exclusions are acceptable only if they occur because of operational or resource

⁴ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

⁵ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

considerations such as excluding persons in hard-to-reach areas. The Consortium asked that each country identify to the extent possible exclusions before sample selection. Adjustments for any noncoverage of the target population in each country was made through benchmarking during the weighting process (see Chapter 15). A complete list of exclusions for countries using population registries is presented in Table 14-6; Table 14-7 includes a similar list for countries not using population registries.

In addition to PIAAC eligible persons not included in sampling frames, persons that were included in the frame but in practice were impossible to be interviewed were treated as exclusions conditional on the total exclusion rate staying at or below 5%. Chapter 16 provides more information about this group, with Table 16-2 showing the overall exclusion rate for each country.

Table 14-6: Portion of target population not covered by Main Study sampling frames for countries using population registries

Country	Percentage of target population not covered*	Group not covered
Austria	0.6%	Undocumented immigrants
Denmark	< 0.1%	Undocumented immigrants
Estonia	2.8%+	Persons without a detailed address; undocumented immigrants (no estimate provided)
Finland	0.2%	Undocumented immigrants; asylum seekers
Flanders (Belgium)	1.0%	Undocumented immigrants
Germany	0.5%	Undocumented immigrants
Italy	0.8%+	Adults in noninstitutional group quarters; undocumented immigrants (no estimate provided)
Japan	2.2%	Nonnationals; undocumented immigrants
Netherlands	0.9%	Undocumented immigrants
Norway	0.4%	Undocumented immigrants
Poland	0.8%	Foreigners staying in Poland fewer than 3 months; nonregistered immigrants
Slovak Republic	0.1%	Undocumented immigrants
Spain	0.0%	None
Sweden	< 1.0%	Undocumented immigrants

* The noncoverage rate accounts for excluded subpopulations such as undocumented immigrants or noninstitutionalized collective DUs, with the exception that the homeless are not being considered part of this rate. Other exclusions that will occur as a natural part of the survey process are not included in the expected noncoverage rate.

Table 14-7: Portion of target population not covered by Main Study sampling frames for countries not using population registries

Country	Percentage of target population not covered*	Group not covered
Australia	3.3%	Persons living in very remote areas, discrete indigenous communities (DIC), or noninstitutional special dwellings; non-Australian diplomats, their staff and household members of such; members (and their dependents) of non-Australian defense forces
Canada	1.8%	Residents of smallest communities in the northern territories; residents of remote and very low population density areas in provinces; and persons living in noninstitutional collective dwellings, other than students in residences.
Cyprus ⁶	< 2.0%	Persons living in houses built after December 2010
Czech Republic	1.8%	Professional armed forces; municipalities with < 200 inhabitants
England/Northern Ireland (UK)	2.0%	Individuals living in private residences that are not listed on the “residential” version of the Postal Address File (PAF) or, in Northern Ireland (UK), not listed on the NI(POINTER) database
France	< 2.6%	Young adults who have never claimed any income and are not attached to their parents households; undocumented immigrants
Ireland	0.4%	Some mobile dwellings
Korea	2.4%	Small islands residents
Russian Federation ⁷	1.5%	Chechnya region
United States	0.1%	People in large gated communities

* The noncoverage rate accounts for excluded subpopulations such as undocumented immigrants or noninstitutionalized collective DUs, with the exception that the homeless are not being considered part of this rate. Other exclusions that will occur as a natural part of the survey process are not included in the expected noncoverage rate.

14.3 National sample designs

The PIAAC standard sample design is a self-weighting design of persons (or of households, for countries without person registries). A self-weighting design is achieved when each sample person (or household, if sampling dwelling units) has an equal probability of selection (standard 4.4.3). For countries that are geographically large, the typical sample design is a stratified multistage clustered area sample. For participating countries that are geographically small, the sample design had less clustering and fewer stages of sampling. Also, several countries had lists of households or persons already available from national registries or registries managed by municipalities.

⁶ Please refer to notes A and B regarding Cyprus in the Note to Readers section of this report.

⁷ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

The TSG allow each country to choose a sample design and selection approach that is most optimal and cost effective as long as the sample design applies full selection probability methods. Each participating country was required to produce a probability-based sample, representative of the target population of the country. The PIAAC standards require probability-based samples because they are essential for two main reasons. First, probability sampling encompasses a set of designs that leads to a variety of unbiased sampling approaches that allow analysts to generalize the results to the target population. Second, measures of precision related to survey estimates (i.e., standard errors, margins of error, confidence intervals) can be computed under a probability design only. Hence, statistical tests for differences between survey estimates are possible only under a probability-based design.

The PIAAC standard probabilities of selection as applied to each country's design are presented in section 14.3.1. Section 14.4.1 presents the sample units selected at each stage of selection, while section 14.4.2 presents the sample selection methods. The factors contributing to the sample size determination in each country, and the sample sizes, are presented in section 14.5.

14.3.1 Probabilities of selection based on PIAAC standard design

Each person in the PIAAC target population must have a nonzero probability of selection resulting from the application of established and professionally recognized principles of scientific sampling (standard 4.4.1). As the ultimate sampling unit, each person in the PIAAC target population must have a calculable nonzero probability of selection. That is, every in-scope person must have a chance of being selected into the PIAAC sample. The following presents the PIAAC approach that was recommended for selecting the ultimate sampling unit for one-, two-, three-, and four-stage sample designs, respectively. The approach is based on PIAAC standards and guidelines. Countries were sent the formulas prior to their sample selection process, and they were asked to confirm or to provide formulas showing their deviations from the self-weighting design. The Consortium conducted checks during and after sample selection. Some countries deviated from these formulae due to oversampling (as given in Table 14-2) or alternative sampling formulas. Table 16-8 provides the variation of the base weights, which identifies the countries that achieved self-weighting or near self-weighting designs (a coefficient of variation of less than 0.05). Among the 14 registry countries, self-weighting or near self-weighting designs were achieved by Austria, Flanders (Belgium), Estonia, Finland, Japan, Netherlands, Norway, Slovak Republic and Sweden. Among the nine screener countries (treating England and Northern Ireland as separate designs), self-weighting or near self-weighting of dwelling units was achieved by Cyprus⁸ and the United States.

One-stage sample designs

For a one-stage sample design without any explicit stratification, let

n = total number of persons to be sample, and

N = total number of eligible persons.

The probability of selecting person l is $r = n/N$.

⁸ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Austria was the only country that adapted a one-stage sample design with no explicit stratification.

For a one-stage stratified sample design, let

n_h = number of persons to be sampled in stratum h ; and

N_h = number of eligible persons in stratum h .

Further, let $r = n/N$, then the probability of selecting person l in strata h is

$$P_{hl} = r.$$

The sample size is allocated to strata as

$$n_h = P_{hl} \times N_h = r \times N_h.$$

Seven countries used a one-stage stratified sample design: Flanders (Belgium), Denmark, Estonia, Finland, Netherlands, Norway and Sweden.

Two-stage stratified probability proportionate to size designs

The formulae for the standard PIAAC selection probabilities for each stage are given below.

For the first-stage sample of primary sampling units (PSUs) in the remaining countries, let

m_h = number of PSUs to be sampled in stratum h ;

MOS_{hi} = measure of size for PSU i in stratum h ; and

I_{psu}^h = sampling interval for the selection of PSUs in stratum h .

The probability of selecting PSU i in stratum h is

$$P_{hi} = \frac{m_h \times MOS_{hi}}{\sum_{i \in h} MOS_{hi}} = \frac{MOS_{hi}}{I_{psu}^h}$$

For the second-stage sample of persons, let

n = total number of persons to be sampled;

N = total number of eligible persons;

n_{hi} = number of persons to be sampled in PSU i of stratum h ; and

N_{hi} = number of eligible persons in PSU i of stratum h .

Let $r = n/N$, then the *conditional* probability of selecting person l in PSU i of stratum h is

$$CP_{hil} = \frac{r}{P_{hi}} = r \times \frac{I_{psu}^h}{MOS_{hi}}$$

The *overall* probability of selecting person l in PSU i of stratum h is

$$P_{hil} = P_{hi} \times CP_{hil} = r.$$

The sample size in PSU i of stratum h is

$$n_{hi} = CP_{hil} \times N_{hi} = r \times \frac{\sum_{i \in h} MOS_{hi}}{m_h} \times \frac{N_{hi}}{MOS_{hi}} = r \times I_{psu}^h \times \frac{N_{hi}}{MOS_{hi}}$$

Seven countries used a two-stage stratified sample design: Cyprus,⁹ France, Germany, Japan, Poland, Slovak Republic and Spain. Poland's weights varied due to oversampling and by applying an alternative design implementation strategy. France used a different approach that followed balance sampling (Deville & Tillé, 2004 and Tillé, 2006) that resulted in varying base weights. Germany's design included deep stratification in the context of Cox (1987) and included simulated values for probabilities of selection due to a sampling-related problem. Spain's weights varied due to applying an alternative design implementation strategy.

Three-stage stratified probability proportionate to size (PPS) designs

In a three-stage stratified PPS design, PSUs are selected with a probability proportionate to a measure of size as described below.

For PSU selection in the training countries, let

- m_h = number of PSUs to be sampled in stratum h ;
- MOS_{hi} = measure of size for PSU i in stratum h ; and
- I_{psu}^h = sampling interval for the selection of PSUs in stratum h .

The probability of selecting PSU i in stratum h is

$$P_{hi} = \frac{m_h \times MOS_{hi}}{\sum_{i \in h} MOS_{hi}} = \frac{MOS_{hi}}{I_{psu}^h}$$

For the second stage sample of dwelling units (DUs), let

- d = total number of housing units to be sampled;
- D = total number of housing units in the sampling frame;
- d_{hi} = number of housing units to be sampled in PSU i of stratum h ; and
- D_{hi} = number of housing units in PSU i of stratum h .

Let $r = d/D$, then the *conditional* probability of selecting housing unit k from PSU i in stratum h is

⁹ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

$$CP_{hik} = \frac{r}{P_{hi}} = r \times \frac{I_{psu}^h}{MOS_{hi}}$$

The *overall* probability of selecting housing unit k in PSU i of stratum h is

$$P_{hik} = P_{hi} \times CP_{hik}$$

The DU sample size in a PSU is

$$d_{hi} = CP_{hik} \times D_{hi} = r \times \frac{\sum_{i \in h} MOS_{hi}}{m_h} \times \frac{D_{hi}}{MOS_{hi}} = r \times I_{psu}^h \times \frac{D_{hi}}{MOS_{hi}}$$

For person selection, let

n_{hik} = number of persons to be sampled from housing unit k in PSU i of stratum h ; and

N_{hik} = total number of eligible persons in housing unit k of PSU i in stratum h .

The *conditional* probability of selecting person l from housing unit k in PSU i of stratum h is

$$CP_{hikl} = \frac{n_{hik}}{N_{hik}}$$

The *overall* probability of selecting person l in housing unit k of PSU i of stratum h is

$$P_{hikl} = P_{hi} \times CP_{hik} \times CP_{hikl} = r \times \frac{n_{hik}}{N_{hik}}$$

Canada, Ireland, Italy, Korea and the Northern Ireland design stratum of the United Kingdom all used a three-stage stratified PPS design. Canada's weights varied due to oversampling. Ireland implemented a sample size-based design in lieu of rate-based design, which caused some variation in the base weights. Italy, Korea and Northern Ireland (UK) each applied an alternative design implementation strategy that caused variation, excessive in the case of Northern Ireland (UK), in the resulting base weights.

Four-stage stratified probability proportionate to size designs

Within the four-stage stratified PPS sample design, PSUs and secondary selection units (SSUs) are selected with a probability proportionate to a measure of size (MOS) as described below.

For PSU selection in the remaining countries, let

m_h = number of PSUs to be sampled in stratum h ; and

MOS_{hi} = measure of size for PSU i in stratum h .

The probability of selecting PSU i in stratum h is

$$P_{hi} = \frac{m_h \times MOS_{hi}}{\sum_{i \in h} MOS_{hi}}$$

For SSU selection, let

q = total number of SSUs to be sampled;

MOS_{hij} = measure of size for SSU j of PSU i in stratum h ; and

I_{SSU} = sampling interval for the selection of SSUs.

The *conditional* probability of selecting SSU j from PSU i in stratum h is

$$CP_{hij} = \frac{q \times \left(\frac{MOS_{hij}}{P_{hi}}\right)}{\sum_{hij} \left(\frac{MOS_{hij}}{P_{hi}}\right)} = \frac{MOS_{hij}/P_{hi}}{I_{SSU}}$$

For DU selection, let

d = total number of housing units to be sampled;

D = total number of housing units in the sampling frame;

d_{hij} = number of housing units to be sampled in SSU j of PSU i of stratum h ; and

D_{hij} = number of housing units in SSU j of PSU i of stratum h .

Let $r = d/D$, then the *conditional* probability of selecting housing unit k from SSU j of PSU i in stratum h is

$$CP_{hijk} = \frac{r}{P_{hi} \times CP_{hij}} = \frac{r \times I_{SSU}}{MOS_{hij}}$$

The *overall* probability of selecting housing unit k in SSU j of PSU i of stratum h is

$$P_{hijk} = P_{hi} \times CP_{hij} \times CP_{hijk} = r$$

The DU sample size in a SSU is

$$d_{hij} = CP_{hijk} \times D_{hij} = r \times I_{SSU} \times \frac{D_{hij}}{MOS_{hij}}$$

For person selection, let

n_{hijk} = number of persons to be sampled from housing unit k of SSU j in PSU i within stratum h ; and

N_{hijk} = total number of eligible persons in housing unit k of SSU j in PSU i within stratum h .

The *conditional* probability of selecting person l from housing unit k of SSU j in PSU i within stratum h is

$$CP_{hijkl} = \frac{n_{hijk}}{N_{hijk}}$$

The *overall* probability of selecting person l from housing unit k of SSU j in PSU i within stratum h is

$$P_{hijkl} = P_{hi} \times CP_{hij} \times CP_{hijk} \times CP_{hijkl} = r \times \frac{n_{hijk}}{N_{hijk}}$$

Australia, the Czech Republic, the Russian Federation,¹⁰ the England design stratum of the United Kingdom, and the United States used a four-stage stratified PPS sample design. The Czech Republic conducted oversampling and also implemented a sequential selection design strategy that caused excessive variation in the resulting base weights. England (UK) had variation in its base weights due to implementing a selection process that is different from the one outlined with the above formulae.

14.4 Sample units and sample selection methods

14.4.1 Sample units

Because Austria, Flanders (Belgium), Denmark, Estonia, Finland, Netherlands, Norway and Sweden all implemented a one-stage sample design, they have only one sample unit: persons. The sampling units for countries with two-, three-, and four-stage sample designs are shown in Tables 14-8 to 14-10, respectively.

Table 14-8: Main study sample units for countries with two stages of sampling

Country		Stage 1	Stage 2
Cyprus ¹¹		Households	Persons
France		Area PSUs	Persons
Germany		Communities	Persons
Japan		Cho/Chome/Aza administrative districts	Persons
Poland	Urban	Towns/Cities	Persons
	Rural	Towns/Villages	Persons
Slovak Republic		Municipalities	Persons
Spain		Area PSUs	Persons

Note: "Area PSUs" indicates primary sampling unit covers a geographic area not defined by a generic geographic terminology (towns, villages, etc).

¹⁰ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

¹¹ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 14-9: Main study sample units for countries with three stages of sampling

Country	Stage 1	Stage 2	Stage 3
Canada	Area PSUs	DUs	Persons
Ireland	Area PSUs	Households	Persons
Italy	Municipalities	Households	Persons
Korea	Enumeration districts	DUs	Persons

Note: “Area PSUs” indicates primary unit covers a geographic area not defined by a generic geographic terminology (towns, villages, etc).

“DUs” indicates dwelling units; “Households” are occupied DUs.

Table 14-10: Main Study sample units for countries with four stages of sampling

Country	Stage 1	Stage 2	Stage 3	Stage 4
Australia	Area PSUs	Blocks	DUs	Persons
Czech Republic	Districts (sub-regions)	Streets	DUs	Persons
England/Northern Ireland (UK)	Postal sectors Addresses	Addresses Households	Households Persons	Persons
Russian Federation ¹²	Regions	Settlements	DUs	Persons
United States	Area PSUs	Area SSUs	DUs	Persons

Note: “Area PSUs” or “Area SSUs” indicates primary or secondary sampling unit covers a geographic area not defined by a generic geographic terminology (towns, villages, etc).

“DUs” indicates dwelling units; “Households” are occupied DUs.

14.4.2 Sample selection methods

Details regarding the selection methods for countries with one- or two-stage sample designs are presented in Tables 14-11 and 14-12, respectively.

Table 14-11: Main Study selection methods for countries with one stage of selection

Country	Description
Austria	Systematic random sample from a sorted list
Denmark	SRS within explicit strata
Estonia	Systematic random from a sorted list within explicit strata
Finland	Systematic random from a sorted list within explicit strata
Flanders (Belgium)	Systematic random from a sorted list within explicit strata
Netherlands	SRS within explicit strata
Norway	SRS within explicit strata
Sweden	SRS within explicit strata

Note: “SRS” indicates simple random sampling.

¹² Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 14-12: Main Study selection methods for countries with two stages of selection

Country		Stage	Description
Cyprus ¹³		1	Systematic random from a sorted list within explicit strata
		2	SRS of 1 person per household via pre-assigned selection grid
France		1	Systematic random from master sample IAAs (master sample selected using the balanced sampling algorithm, the “Cube” method, PPS (number of main residences in the IAA))
		2	Systematic random from a sorted list
Germany		1	Stratified, PPS (target population) with allocation by controlled rounding
		2	Two-phase sample. <ul style="list-style-type: none"> Phase 1: The registries of the selected communities were asked to select an EPSEM sample of individuals. Phase 2: Within each community, the individuals selected in Phase 1 were allocated to a matrix that was divided into six age groups x gender. Allocation of the Phase 2 sample size was done using an Iterative Proportional Fitting (IPF) procedure. The selection of persons within a community was done by systematic random sampling with a random start number and a sampling interval.
Japan		1	Systematic PPS (number of inhabitants age 15-64 as of March 2010) from a sorted list within explicit strata
		2	Systematic random from a sorted list
Poland	Urban	1	All towns/cities selected with certainty
		2	SRS within explicit strata
	Rural	1	PPS (population age 16-65) within explicit strata
		2	SRS without replacement of clusters of 8 persons in explicit strata
Slovak Republic		1	Systematic PPS (population age 16-65) from a sorted list within explicit strata
		2	Systematic random from a sorted list
Spain		1	Systematic PPS (population) from a sorted list within explicit strata
		2	Systematic random from a sorted list

Note: “SRS” indicates simple random sampling.

All countries with three- or four-stage designs selected samples of dwelling units before the enumeration and selection of persons within households. Although the goal was to select one person per household, the selection of more than one person per household was preferred for countries with a large variation in household size (standard 4.4.4). These include the Russian Federation¹⁴ and the United States. Details regarding the selection methods for countries with three- or four -stage designs are presented in Tables 14-13 and 14-14, respectively.

¹³ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

¹⁴ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 14-13: Main Study selection methods for countries with three stages of selection

Country	Stage	Description
Canada	1	Systematic PPS (2006 population counts) from a sorted list within explicit strata with Census Metropolitan Areas sampled with certainty
	2	Systematic random from a sorted list within explicit strata
	3	SRS of 1 person per household via pre-assigned hash number
Ireland	1	Stratified PPS (total dwellings)
	2	SRS
	3	SRS of 1 person per household
Italy	1	Systematic PPS (target population) from a sorted list within explicit strata
	2	Systematic random from a sorted list
	3	SRS of 1 person per household via selection grid is used if the household composition is different from the register; otherwise SRS from registry.
Korea	1	Systematic random sample from a sorted list within explicit strata
	2	Systematic random from a sorted list
	3	SRS of 1 person per household

Note: “SRS” indicates simple random sampling.

Table 14-14: Main Study selection methods for countries with four stages of selection

Country	Stage	Description
Australia	1	Systematic PPS (number of DU clusters) from a sorted list within explicit strata (subsample from master sample)
	2	Systematic PPS (number of DU clusters) from a sorted list (subsample from master sample)
	3	Systematic random from a sorted list
	4	SRS of 1 person per household
Czech Republic	1	Systematic PPS (number of inhabitants aged 16-65) from a sorted list within explicit strata
	2	Systematic PPS (number of address points)
	3	SRS; selected a “basic” sample of households to achieve the 5,000 completes plus an additional sample of households in which only 16- to 29-year-olds were sampled.
	4	SRS of 1 person per household
England (UK)	1	Systematic PPS (PAF single occupancy count) from a sorted list within explicit strata
	2	Systematic random from a sorted list
	3	SRS of 1 household at the sampled address using the Kish grid
	4	SRS of 1 person per household using the Kish grid
Northern Ireland (UK)	1	Systematic random from a sorted list
	2	SRS of 1 household at the sampled address using the Kish grid
	3	SRS of 1 person per household using the Kish grid
Russian Federation ¹⁵	1	Systematic PPS (population in the region) from a sorted list within explicit strata
	2	Systematic PPS (target population) from a sorted list
	3	Systematic random from a sorted list
	4	SRS of 1 person for household sizes up to 4 (otherwise 2 persons) via pre-assigned selection grid
United States (USA)	1	Systematic PPS (population) within explicit strata
	2	Systematic PPS (number of DUs) from a sorted list
	3	Systematic random from a sorted list
	4	SRS of 1 person for household size up to 3 (otherwise 2 persons)

Note: “SRS” indicates simple random sampling.

¹⁵ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Stratification combines sample units into homogeneous groups and reduces sampling variability between such groups and thus reduces the overall sampling variance associated with the resulting survey estimates. To maximize the benefit of stratification, stratification variables should be reliable and related to the survey outcome. Many of the countries utilizing population registries have the benefit of person-level characteristics available as stratification variables. The stratification and/or sorting variables for countries with one, two, three, and four stages of selection are detailed in Tables 14-15 to 4-18, respectively.

Table 14-15: Main Study stratification/sorting variables and methods for countries with one stage of selection

Country	Description
Austria	Sort by province, urban/rural, age, gender and citizenship
Denmark	Strata: age categories, immigration status
Estonia	Strata: gender and age categories Within strata: sort by region and age
Finland	Strata: native language (Finnish and other languages than Swedish, and Swedish) Within strata: sort by region, age, educational attainment, and gender
Flanders (Belgium)	Strata: province Within strata: sort by postal code, gender and age
Netherlands	Strata: municipality
Norway	Strata: level of education and age group
Sweden	Strata: gender, age, country of birth, level of education

Table 14-16: Main Study stratification/sorting variables and methods for countries with two stages of selection

Country		Stage	Description
Cyprus ¹⁶		1	Strata: district, urban/rural classification Within strata: sort by geographic location
		2	None
France		1	Strata: administrative region (for master sample) Balancing variables: number of main residences, total income, number of DUs in rural, peri-urban, and urban areas.
		2	Stratified by housing (synthetic variable differentiating ordinary housing and communities) and sorted by department (administrative district).
Germany		1	Strata: region, urban/rural status (BIK) – approximately 1,000 strata cells
		2	None in Phase 1. In Phase 2, stratified by age group and gender, sorted by age.
Japan		1	Strata: region, urban/rural status; Sort by regional code
		2	Sort by address
Poland	Urban	1	Strata: size class
		2	Strata: age (19-26, other)
	Rural	1	Strata: region and size class
		2	Strata: age (19-26, other)
Slovak Republic		1	Strata: region, municipality size; Within strata: sort by number of age 16-65 in municipality
		2	Sort by gender and age
Spain		1	Strata: categories of municipality size Within strata: sort by population size
		2	Sort by gender and age

¹⁶ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 14-17: Main Study stratification/sorting variables and methods for countries with three stages of selection

Country	Stage	Description
Canada	1	Stratify by province, urban/rural; sort by geographic order of PSUs and 2006 population counts
	2	Stratified by province/territory and urban/rural. Sort by geographic order (province/territory code, urban/rural, PSU ID, Census collection unit ID)
	3	None
Ireland	1	Strata: urban/rural status, and educational profile Within strata: sort by size of SAs
	2	None
	3	None
Italy	1	Strata: geographic regions of equal size Within strata: sort by the target population count of the PSUs
	2	None
	3	Random sort if selection from registry. If the household composition is different from the registry, persons are sorted by gender and age and the selection grid is used.
Korea	1	Strata: administrative districts Within strata: sort by enumeration district characteristics, such as townhouse versus apartment, percentage of 1-person household, education level, average age, percentage of people who are older than 60
	2	Sort by address
	3	None

Table 14-18: Main Study stratification/sorting variables and methods for countries with four stages of selection

Country	Stage	Description
Australia	1	Strata: state, part of state Within strata: serpentine sort by geography
	2	Serpentine sort by geography
	3	Serpentine sort by geography
	4	None
Czech Republic	1	Strata: region, municipality size Within strata: sort by code of location
	2	Sort by code of the street
	3	None
	4	Sort by year of birth
England (UK)	1	Strata: region, percentage living in social housing Within strata: sort by percentage of White British
	2	Sort by postcode and address number
	3	Sort by addresses (alphanumerically)
	4	Sort by first name
Northern Ireland (UK)	1	Sort by council ward, postcode within ward, and then alphanumerically within postcode
	2	Sort by addresses (alphanumerically)
	3	Sort by first name
Russian Federation ¹⁷	1	Strata: macro regions Sort by federal county, population size for noncertainty PSUs
	2	Sort by type of settlement
	3	Sort by type of urban district (central/middle/outskirt)
	4	None
United States	1	Strata: region, metro area classification, race/ethnicity, income, percentage of the population that is foreign born
	2	Sort by geographic location
	3	Sort by geographic location
	4	None

14.5 Sample size determination

Adequate sample sizes are needed to establish stable item characteristics and to estimate separate population models for each tested language in a participating country. Population modeling is a critical step in obtaining appropriate proficiency values to be used in describing the distributions of skills in a country and in reporting national and subpopulation data.

The overall goal of the sample design for the Main Study was to obtain a nationally representative sample of the target population in each participating country that is proportional to the population across the country (i.e., a self-weighting sample design). As mentioned earlier, countries had the option of increasing sample sizes to obtain reliable estimates for groups of special interest (e.g., 16- to 29-year-olds), for geographic regions (e.g., states and provinces) or to extend the age range (e.g., 66-plus). However, the minimum sample size required was for a

¹⁷ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

self-weighting design, and any sample size attributable to oversampling, or to subgroups outside of the PIAAC target population, was additional. PIAAC target sample sizes are presented in section 14.5.1.

To determine the initial sample size for the Main Study, the required number of assessments had to be adjusted to account for survey ineligibility and expected nonresponse to both the BQ/JRA and the assessment. For countries with a household screener, sample size goals had to be constructed for the screener to account for ineligibility and screener nonresponse, in addition to nonresponse to the BQ/JRA and assessment.

In most highly clustered surveys or those with a high degree of variability in sampling rates due to oversampling, initial sample sizes must be increased to retain the desired precision. For PIAAC, countries were asked to estimate the design effect of their design with such an increase in mind (guideline 4.3.2.B). However, the guideline was relaxed for this first cycle of PIAAC due to (1) uncertainties surrounding the quality of the design effect estimates produced using the Field Test data and (2) the limited amount of time available between the Field Test and the Main Study to allow changes to sample size goals of the survey.

Instead, countries with estimated large design effects were asked to modify their design to the extent possible to reduce the clustering of the sample. To compute the initial sample size, countries were allowed to use a design effect of 1.50 (if the expected design effect was greater than 1.50). However, countries are asked to report their best estimate of the design effect so that improvements to clustering and stratification may be identified for future cycles of PIAAC.

Section 14.5.2 contains information about the various expected eligibility rates used in the computation of the initial sample sizes by the participating countries and the plans for selecting reserve samples in case observed rates were different from the expected ones.

14.5.1 PIAAC target sample sizes

The minimum sample size requirements for the Main Study for the standard target population speaking the main language of the country was dependent on the optional components of the psychometric assessments administered in the country:

- Both problem solving and reading components – 5,000 minimum completes
- Problem solving only – 5,000 minimum completes
- Reading only – 4,500 minimum completes
- No optional components – 4,500 minimum completes

The definition of a completed case is given in TSG 4.3.3 as follows:

‘Standard 4.3.3 A completed case is one that contains at least the following:

- *Responses to key background questions, including age, gender, highest level of schooling and employment status; and*
- *A completed Core instrument (i.e. the interviewer asked the respondent all Core questions or the Core instrument was not completed for a literacy-related reason [e.g. because of a language difficulty] or because the respondent was unable to read or write in any of a country’s PIAAC official languages); or*

- *Responses to age and gender for literacy-related nonrespondents to the BQ/JRA.’*

To obtain a self-weighting standard design, the number of assessments in any other language had to be proportional to the number of people speaking the additional languages in the country. Countries that planned to report on general proficiency, regardless of the languages tested, had to achieve the appropriate minimum completed sample size shown above for their main language. Thus, the minimum sample size requirement for an individual country not only depended on the optional psychometric assessments administered and the number of languages being tested but also the number of reporting languages determined by the country.

Most countries conducted both the reading and problem-solving components. Cyprus,¹⁸ Italy and Spain conducted the reading components only; Finland, Japan and the Russian Federation¹⁹ conducted the problem-solving component only. France declined both optional assessments. Five countries performed the assessment in multiple languages. Canada, Estonia, Finland and the Slovak Republic conducted assessments in two languages; Spain conducted the assessment in five languages. The full list of the optional components of the psychometric assessment being conducted by the countries, including the languages of the assessments and the resulting required number of assessments, is presented in Table 14-19, and target sample sizes are given in Table 14-20 below.

¹⁸ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

¹⁹ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 14-19: Required sample sizes by assessment language

Country	Assessment language and proportion of population speaking it (as available)	Optional components of psychometric assessment being conducted	Required sample size (general proficiency reporting in terms of language unless otherwise indicated) ¹
Australia	English	R, PS	5,000
Austria	German (88.5%)	R, PS	5,000
Canada	Canadian English (67.3%)	R, PS	5,000
	French (21.1)	R, PS	5,000
Cyprus ²⁰	Greek (84.1%)	R	4,500
Czech Republic	Czech	R, PS	5,000
Denmark	Danish (92%)	R, PS	5,000
England/N. Ireland (UK)	UK English	R, PS	5,000
	UK English	R, PS	5,000
Estonia	Estonian (67%)	R, PS	5,000
	Russian (33%)	R, PS	2,500
Finland	Finnish (90.5%)	PS	5,000
	Swedish (5%)	PS	276
Flanders (Belgium)	Dutch	R, PS	5,000
France	French	None	4,500
Germany	German	R, PS	5,000
Ireland	English	R, PS	5,000
Italy	Italian	R	4 500
Japan	Japanese (~100%)	PS	5,000
Korea	Korean	R, PS	5,000
Netherlands	Dutch	R, PS	5,000
Norway	Norwegian (Bokmål)	R, PS	5,000
Poland	Polish	R, PS	5,000
Russian Federation ²¹	Russian (98.2%)	PS	5,000
Slovak Republic	Slovak (89.8%)	R, PS	5,000
	Hungarian (10.2%)	R, PS	568
Spain	Castellano (60%)	R	4,500
	Gallego (6%)	R	225
	Catalan (18%)	R	675
	Valencian (11%)	R	410
	Euskera (5%)	R	190
Sweden	Swedish	R, PS	5,000
United States	English (91.5%)	R, PS	5,000

¹ The required sample size in this table does not consider the occurrence of oversampling in some countries.

²⁰ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

²¹ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

14.5.2 Eligibility rates and reserve samples

The eligibility rate assumptions specified by countries were reviewed to help ensure that initial sample sizes were large enough to achieve the required number of assessments. Countries including a dwelling unit sample as part of their sample design were further required to provide an estimated screener eligibility rate. Selected units found to be vacant, for seasonal use only, not actually dwelling units, or without persons ages 16 to 65 were considered ineligible for the survey and had to be accounted for in the derivation of the final sample size.

The expected response rates reported during the National Survey Design and Planning Report process were taken into account to ensure that the initial samples sizes were large enough to yield the required number of assessments. Some adjustments to these expected rates were made based on Field Test experience.

It is difficult to predict the nonresponse and ineligibility rates for a survey like PIAAC. As a result, the Consortium encouraged each country to consider selecting a reserve sample of 10% or more of the size of the main initial (original) sample. The requirement was to select the reserve sample at the same time as the original sample and then set it aside and not use it unless sample monitoring showed potential for shortfall. Reserve samples were recommended over supplemental samples because computing the selection probabilities is simpler with a reserve sample than supplemental samples. The same concept was used if a country was concerned about exceeding the target sample size by a significant amount. After selecting a 110% sample, the country was able to release to the field a sample that was less than 100% by randomly selecting (subsetting) from the original sample and then releasing more sample as needed. Also the countries could split the reserve sample randomly into several “release” groups as long as the release group by itself was representative of the country (not any particular subgroup).

The target sample sizes for each stage, including the target person sample sizes, are presented in Table 14-20.

Table 14-20: Main Study target sample sizes

Country	Sample size				Target number of completes*	PIAAC standard**
	PSUs	SSUs	DUs	Persons		
Australia	2,136	2,136	14,423	11,250	9,000 ¹	5,000
Austria				10,000	5,000	5,000
Canada ²	217		49,234	34,464	25,267	10,000
Cyprus ²²			16,215	4,986	4,500	4,500
Czech Republic ³	284	400	15,660	6,312	6,000	5,000
Denmark ⁴				14 100	6 900	5,000
England (UK)	488	13,664	13,664	7,429	4,850	5,000
Estonia				13,000	7,500	7,500
Finland				8,000	5,300	5,276
Flanders (Belgium)				10,960	5,000	5,000
France	525			10,500	5,200	4,500
Germany	320			11,406	5,000	5,000
Ireland	700		13,600	8 092	6,200	5,000
Italy	260		17,520	7,742	4,500	4,500
Japan	459			13,000	5,000	5,000
Korea	883		8,330	7,296	5,000	5,000
Netherlands				10,256	5,000	5,000
Norway				9,453	5,000	5,000
Northern Ireland (UK)		9,470	9,470	5,143	3,492	5,000
Poland	85 urban 1,086 rural			13,430	9,132 ⁵	5,000
Russian Federation ²³	25 ⁶	93	9,630	5,540	5,000	5,000
Slovak Republic	562			9,280	5,568	5,568
Spain	1,200			14,400	6,000	6,000
Sweden				10,000	5,100	5,000
United States	80	901	9,610	6,371	5,000	5,000

■ indicates there is no such stage in the country's sample design.

*Targets include multiple languages and oversampling within target population, unless otherwise noted.

** Targets include multiple languages; there are no PIAAC standards for oversampled populations.

¹ 7,922 of the targeted completes were expected to be ages 16-65.

² Values include oversamples of 20,488 dwellings and 14,342 persons for 9,756 completes.

³ Values include 5,923 sampled DUs, 1,052 sampled persons, and 1,000 targeted completes for the country-specific sample.

⁴ Values do not include the Programme for International Student Assessment oversample, which was not part of the PIAAC sample.

⁵ Includes oversample of 5,000 persons ages 19-26.

⁶ Although the Russian Federation selected 25 PSUs, only 23 PSUs were included in the final analyses (Moscow and Moscow region were excluded due to data issues)

²² Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

²³ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

14.6 Sample selection results

Table 14-21 provides the final sample sizes for each stage of sampling for each country. Table 16-7 provides the final number of respondents (with a final sampling weight).

Table 14-21: Main Study selected sample sizes by sampling stage

Country	Sample size			
	PSUs	SSUs	DUs	Persons
Australia	~2,200	~2,200	14,634	9,725 ¹
Austria				10,000
Canada	217		49,487	33,987
Cyprus ²⁴			8,514	5,095
Czech Republic	284	400	17,069	6,907
Denmark				16,040
England (UK)	488	13,664	13,664	7,933
Estonia				13,000
Finland				8,099
Flanders (Belgium)				9,200
France	525			10,500
Germany	277			10,240
Ireland	700		10,500	6,442
Italy	260		11,592	7,377
Japan	459			11,000
Korea	883		8,330	7,296
Netherlands				10,256
Northern Ireland (UK)		9,480	9,480	4,937
Norway				8,506
Poland	85 urban 1,086 rural			18 774
Russian Federation ²⁵	25 ²	93	9,376	4,199
Slovak Republic	562			9,280
Spain	1,200			14,400
Sweden				10,000
United States	80	896	9,468	6,100

■ indicates that there is no such stage in the country's sample design.

¹ 8,433 were ages 16-65.

² Although the Russian Federation selected 25 PSUs, only 23 PSUs were included in the final analyses (Moscow and Moscow region were excluded due to data issues)

14.7 Sampling quality control checks

The Consortium developed a comprehensive set of quality assurance and quality control checks to ensure PIAAC produced high-quality data that were comparable across countries. Section 16.1 contains a description of the quality assurance and quality control procedures developed for all sampling activities, including sample design and selection results. Countries were required to

²⁴ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

²⁵ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

complete quality control sample selection forms, which collected sampling information for each stage of selection using standard templates. The templates were designed to capture aggregated information that was necessary for verifying that the sample was representative of the target population and that sampling was conducted in an unbiased and randomized way. For example, at each stage countries were asked to estimate and report the total target population within each stratum so that distributions by stratum could be reviewed at each sampling stage. The Consortium carried out all sampling quality control checks as listed in section 16.1 and informed the countries of the approval of their plans/procedures or asked for revisions to aspects that did not meet the PIAAC standards.

Table 14-22 provides a summary of the sample design and selection quality assessment. For the sampling plan, it was essential that a complete sampling plan was provided, and that the country responded to feedback from the Consortium. For the sampling plan, a cautionary remark was given to the Russian Federation²⁶ due to an insufficient number of PSUs selected. As it relates to the sample selection process conducted in the country's home office, it was important that complete QC sample selection forms were provided prior to data collection, that each person in the PIAAC target population had a nonzero and known (calculable) probability of selection resulting from the application of established and professionally recognized principles of scientific sampling, and that there was no substitution of sampling units. As indicated in Table 14-22, cautionary remarks were given to Australia (quality level unknown due to country confidentiality restrictions or unavailability of data), Czech Republic (for late sample selection forms), Germany (for simulated probabilities of selection), the Russian Federation²⁷ (noncompliance in completing the quality control forms), and Japan (for an approved deviation of the TSG, given the disastrous earthquake. The design accounted for the affected PSUs through combining strata, increasing sample sizes in affected strata, and using weighting procedures to reduce bias), With regard to sample selection processes that were conducted in the field, countries were assessed according to the following criteria ensuring that:

- persons were selected from within households using a fully enumerated grid of household members,
- each person in the PIAAC target population had a nonzero and known (calculable) probability of selection resulting from the application of established and professionally recognized principles of scientific sampling,
- no more than two persons were selected in a household,
- less than 10% of households had two persons selected, and
- there was no substitution of sampling units.

Only cautionary remarks were given to Australia (quality level unknown due to country confidentiality restrictions or unavailability of data) and the UK (imputed theoretical person base weights for 52 cases (49 in England and three in Northern Ireland) due to a technical problem with the contact data that the interviewers entered).

²⁶ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

²⁷ Please refer to the above note regarding the Russian Federation.

Table 14-22: PIAAC sample design and selection outcome summary

	Sampling Plan	Sample Selection	
		Home Office	In Field
Australia	P	C-U	C-U
Austria	P	P	N/A
Flanders (Belgium)	P	P	N/A
Canada	P	P	P
Cyprus ²⁸	P	P	P
Czech Republic	P	C-NC	P
Denmark	P	P	N/A
England (UK)	P	P	C-PC
Estonia	P	P	N/A
Finland	P	P	N/A
Germany	P	C	N/A
Ireland	P	P	P
Italy	P	P	P
Japan	P	C-A	N/A
Korea	P	P	P
Netherlands	P	P	N/A
Northern Ireland (UK)	P	P	C-PC
Norway	P	P	N/A
Poland	P	P	N/A
Russian Federation ²⁹	C-PC	C-NC	P
Slovak Republic	P	P	N/A
Spain	P	P	N/A
Sweden	P	P	N/A
United States	P	P	P

P: Pass (relevant requirement completely met)

C: Caution (relevant requirement met to a reasonable extent)

C-A: Caution, approved deviation

C-NC: Caution, did not comply

C-PC: Caution, partial compliance

C-U: Caution, quality level unknown due to country confidentiality restrictions or unavailability of data

14.8 Respondent incentives

Respondent incentives have been shown to be effective for improving response rates without affecting the respondent's performance. As a result, the use of incentives can potentially reduce bias in the estimates. As such, countries were permitted to offer modest incentives to obtain respondent cooperation, such as a monetary or nonmonetary incentive (e.g., pen, notepad, candy, mug, voucher, gift certificate). A variety of incentives were offered across the participating countries with the exception of two countries: Australia and Canada have rules preventing the

²⁸ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

²⁹ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

use of incentives in government surveys. Table 6-8 contains the type of incentives used during the Main Study data collection in PIAAC.

14.9 Recommendations for future cycles

Based on the Field Test and Main Study experience of PIAAC Round 1, the Consortium is proposing a series of recommendations for future cycles of PIAAC as it relates to sampling activities.

1. Countries should follow the TSG on the qualifications of the National Sampling Manager.
2. The Consortium and countries should work together to provide the BQ in as many languages as possible so that background information can be used in the generation of plausible values in case the person speaks a different language than the assessment language(s) offered.
3. Countries should evaluate the quality of the frames from the start so they have adequate time to look for alternatives if the quality (and coverage) of the frame does not meet the standards.
4. Before countries move forward with the sample that has been selected, the QC sample selection forms must be reviewed by the Consortium, with feedback provided.
5. Before countries submit sample monitoring forms, all numbers should be double checked. The Consortium will insert some automated checks into the forms to help ensure the forms are completed accurately.
6. Countries should use the Response Rate Toolkit to compute the response rates for the forms, or to check any automated program that was developed.
7. Countries should use the results of PIAAC to improve upon the stratification and sorting scheme. The nonresponse bias analysis and the scores can be used to identify better stratification and sorting variables, such as education, employment and other variables that are correlated with the scores.
8. Countries should use the design effects to identify ways to improve the sample design. That is, countries should evaluate how to reduce the clustering and unequal probabilities effects as plans occur for the next cycle.
9. While preparing plans for the next cycle, initial sample sizes should take into account the impact of the design components (cluster sizes, stratification, variation in weights, multiple imputation) on the resulting DEFFs observed in Cycle 1 (or an expected DEFF due to design improvements since Cycle 1) so that the quality of the resulting estimates is comparable across countries. Countries should plan to increase their sample sizes to account for the large design effects to arrive at an acceptable effective sample size, or make changes in their sample designs to reduce design effects.
10. Countries need to follow the schedules of all QC sampling activities so there is adequate time to identify problems and to incorporate changes to correct mistakes in a timely fashion.

References

- Cox, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82(398), 520-524.
- Deville, J., & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 893-912.
- Tillé, Y. (2006). *Sampling algorithms*. New York, NY: Springer.

Chapter 15: Survey Weighting and Variance Estimation

Leyla Mohadjer, Tom Krenzke, Wendy Van de Kerckhove and Valerie Hsu, Westat

This chapter describes the methods that countries used to compute sampling weights and estimate variances through the use of replicate weights. The purpose of calculating sampling weights for PIAAC is to permit inferences from persons included in the sample to the population from which they were drawn and to have the tabulations reflect estimates of the population totals. Sampling weights can be considered as estimated measures of the number of units in the target population that a sampled case represents. Weighting incorporates several features of the survey, including the probabilities of selection of units in the sample and adjustments for nonresponse and any known differences between the selected sample and the total target population. Differences between the sample and the population may arise because of sampling variability, differential response rates or coverage rates among subgroups of the population, and other types of response errors, such as misclassification errors.

In PIAAC, survey weighting was performed to accomplish the following objectives:

- To permit unbiased estimates by compensating for possible disproportionate sampling of various subgroups in the sample
- To minimize biases arising from differences between respondents and nonrespondents
- To compensate for noncoverage in the sample due to inadequacies in the sampling frame or other reasons for noncoverage
- To bring data up to the dimensions of the population totals
- To reduce sampling errors by using auxiliary data on population characteristics that are known with a high degree of accuracy
- To facilitate the estimation of variances through the use of the replication approach

15.1 Survey weighting

Weighting involves designing adjustment factors to compensate for variable probabilities of selection and to reduce potential bias due to nonresponse, deficiencies in the sampling frame and other complications that may arise during the sample selection process. This section provides a description of the standard weighting steps employed in the first round of PIAAC. Countries were required to follow the weighting process outlined in the PIAAC Weighting and Variance Estimation Plan produced by the Consortium, which followed the standards and guidelines in Section 14 of the PIAAC Technical Standards and Guidelines. It described the weighting process, including the weighting steps, treatment of different disposition codes, calculation of weighting adjustment factors, assignment of variance strata and variance units, and creation of replicate weights. Using the weighting approach described in the Weighting and Variance

Estimation Plan for all countries ensured comparable estimates of proficiency and their sampling error across countries.

A final weight is required for all sampled persons with a completed BQ and BQ literacy-related nonrespondents (LRNRs) with age and gender collected. The BQ LRNRs with age and gender collected receive a final weight despite the lack of BQ or assessment data because they are considered part of the PIAAC target population and cannot be represented by survey respondents (see section 15.1.3). There were a number of steps in the development of the final weights intended for use in the estimation and analysis:

1. Assignment of a household base weight to each sampled household to compensate for differential probabilities of selection (for screener countries¹ only)
2. Household-level eligibility and nonresponse adjustments to reduce potential biases arising from differences between respondents and nonrespondents (for screener countries only)
3. Assignment of a person base weight to each sampled person to compensate for differential probabilities of selection
4. Person-level eligibility adjustment (for registry countries² only) and nonresponse adjustments
5. Trimming to reduce the impact of large weights, if necessary
6. Calibration of the person weights to independent control totals to compensate for noncoverage in the sample due to deficiencies in the sampling frame

The succeeding sections describe each of the weighting steps in detail. A summary of the adjustment factors and resulting weights at each weighting step is provided in Tables 15-1A and 15-1B for registry and screener countries, respectively.

¹ Screener countries refer to countries whose sample design included a screener stage.

² Registry countries refer to countries whose sample design did not include a screener stage.

Table 15-1A: Adjustment factors and weights for registry countries

Weighting Step	Factor	Weight
Base weight	N/A	$W_l = \frac{1}{P_{hl}}$
Unknown eligibility adjustment	$F_{1l} = \begin{cases} \frac{S_R + S_{NR} + S_{L1} + S_{L2} + S_D + S_I + S_U}{S_R + S_{NR} + S_{L1} + S_{L2} + S_D + S_I} & \text{if } l \in I \\ \frac{S_R + S_{NR} + S_{L1} + S_{L2} + S_D}{S_R + S_{NR} + S_{L1} + S_{L2} + S_D + S_I} & \text{if } l \in U \\ 1 & \text{if } l \in R, NR, L1, L2, D \end{cases}$	$W_l F_{1l}$
Nonliteracy-related nonresponse adjustment	$F_{3l} = \begin{cases} 1 & \text{if } l \in L1, L2, I \\ \frac{S_R + S_{NR} + S_D + S_U}{S_R} & \text{if } l \in R \\ 0 & \text{if } l \in NR, D, U \end{cases}$	$W_l F_{1l} F_{3l}$
Literacy-related nonresponse adjustment	$F_{4l} = \begin{cases} 1 & \text{if } l \in R, I \\ \frac{S_{L1} + S_{L2}}{S_{L1}} & \text{if } l \in L1 \\ 0 & \text{if } l \in L2 \end{cases}$	$W_l F_{1l} F_{3l} F_{4l}$
Trimming*	$F_{5l} = \begin{cases} 1 & \text{if } W_l F_{1l} F_{3l} F_{4l} \leq cutoff \\ \frac{cutoff}{W_l F_{1l} F_{3l} F_{4l}} & \text{if } W_l F_{1l} F_{3l} F_{4l} > cutoff \end{cases}$	$W_l F_{1l} F_{3l} F_{4l} F_{5l}$
Calibration	$F_{6l} = \frac{S^*}{S_R + S_{L1}} \text{ (for post-stratification)}$ <p>See Deming and Stephan (1940) for raking adjustments and Särndal, Swenson, and Wretman (1992) for GREG estimation.</p>	$W_l F_{1l} F_{3l} F_{4l} F_{5l} F_{6l}$

* If the Consortium computed the sampling weights, an initial calibration step was performed prior to trimming (i.e., one iteration of calibration, trimming (if necessary), and recalibration was performed following the nonresponse adjustments).

Note: The factors and weights shown here are for a person l . The persons can be classified as R: BQ respondent who is not assessment literacy-related nonrespondent, L1: BQ literacy-related nonrespondent with age and gender successfully collected or assessment literacy-related nonrespondent, L2: BQ literacy-related nonrespondent with age or gender not successfully collected, NR: BQ nonliteracy-related nonrespondent, I: ineligible, D: sampled person with a disability, or U: sampled person with unknown eligibility status. S represents the sum of the prior-stage weights over records in the same adjustment cell as person l , and S^* is the control total for the cell. P represents the selection probability. The factor F_2 is reserved for countries with screeners.

Table 15-1B: Adjustment factors and weights for screener countries

Stage	Weighting Step	Factor	Weight
Screener	Base weight	N/A	$W_k = \frac{1}{P_{hi} CP_{hik}}$
	Unknown eligibility a $W_l F_{3l} F_{4l} F_{5l} F_{6l}$ djustment	$F_{1k} = \begin{cases} \frac{S_L + S_R + S_{NR} + S_I + S_U}{S_L + S_R + S_{NR} + S_I} & \text{if } k \in I \\ \frac{S_L + S_R + S_{NR}}{S_L + S_R + S_{NR} + S_I} & \text{if } k \in U \\ 1 & \text{if } k \in L, R, NR \end{cases}$	$W_k F_{1k}$
	Nonresponse adjustment	$F_{2k} = \begin{cases} 1 & \text{if } k \in L, I \\ \frac{S_R + S_{NR} + S_U}{S_R} & \text{if } k \in R \\ 0 & \text{if } k \in NR, U \end{cases}$	$W_k F_{1k} F_{2k}$
BQ	Base weight	N/A	$W_l = W_k F_{1k} F_{2k} \frac{1}{CP_h}$
	Nonliteracy-related Nonresponse adjustment	$F_{3l} = \begin{cases} 1 & \text{if } l \in L, I \\ \frac{S_R + S_{NR} + S_D}{S_R} & \text{if } l \in R \\ 0 & \text{if } l \in NR, D \end{cases}$	$W_l F_{3l}$
	Literacy-related nonresponse adjustment	$F_{4l} = \begin{cases} 1 & \text{if } l \notin L \\ \frac{S_L^{BQ} + S_L^{MAIN} + S_L^{SCR}}{S_L^{BQ} + S_L^{MAIN}} & \text{if } l \in L^{BQ} \text{ or } L^{MAIN} \\ 0 & \text{if } l \in L^{SCR} \end{cases}$	$W_l F_{3l} F_{4l}$
	Trimming*	$F_{5l} = \begin{cases} 1 & \text{if } W_l F_{3l} F_{4l} \leq cutoff \\ \frac{cutoff}{W_l F_{3l} F_{4l}} & \text{if } W_l F_{3l} F_{4l} > cutoff \end{cases}$	$W_l F_{3l} F_{4l} F_{5l}$
	Calibration	$F_{6l} = \frac{S^*}{S_R + S_L^{BQ} + S_L^{MAIN}}$ (for post-stratification) See Deming and Stephan (1940) for raking adjustments and Särndal, Swenson, and Wretman (1992) for GREG estimation.	

* If the Consortium computed the sampling weights, an initial calibration step was performed prior to trimming (i.e., one iteration of calibration, trimming (if necessary), and recalibration was performed following the nonresponse adjustments).

NOTE: The factors and weights shown here are for a household k or person l . The households and persons can be classified as R: respondent, L: literacy-related nonrespondent, NR: nonliteracy-related nonrespondent, I: ineligible, D: sampled person with a disability, or U: unknown eligibility. S represents the sum of the prior-stage weights over records in the same adjustment cell as household k or person l , S^* is the sum of screener base weights, and S^* is the control total for the cell. P represents the selection probability.

15.1.1 Preliminary steps in weighting

Countries were responsible for selecting the variables that were used in their nonresponse and calibration weighting adjustments. Prior to weighting, countries were required to evaluate the variables being considered for the weighting adjustments in their PIAAC main sample.

For the nonresponse adjustment, variables needed to be available for all eligible units and be related to proficiency and response propensity. The pool of potential nonresponse adjustment variables came from the sampling frame (and/or the screener) or other external sources. A common source of nonresponse adjustment variables for screener countries was a country census. For registry countries, the registry data were highly beneficial during the nonresponse adjustment.

For the calibration adjustment, all variables selected by countries were required to have reliable control totals and be available for all BQ respondents and LRNRs with age and gender collected. The quality of the data from the external sources had to exceed the quality of data from PIAAC (e.g., the mean square errors of the external estimates needed to be smaller than those of the uncalibrated estimates from the survey). The concepts, definitions and coverage of the data (counts) from the external sources needed to be the same as those employed by PIAAC. Additionally, the year of the control totals needed to be as close to the data collection period as possible, ideally covering the same time period as the field period.

Variables used for nonresponse adjustment and in calibration must have less than 5% missing data. If the amount of missing data of the variables used in weighting adjustments did not exceed the 5% threshold, countries were required to follow the weighting standards and guidelines on imputing for missing data.

15.1.2 Household-level weighting adjustments

This section outlines the weighting process at the household level for screener countries, which included the creation of the household base weights that reflected the household selection probability and was adjusted for unknown eligibility and nonresponse to the screener.

Household base weights

For screener countries, the household base weight was assigned to all sampled households and was computed as the reciprocal of the household selection probability. For screener countries with a multistage sample design, the household selection probability corresponded to the product of the conditional selection probabilities at each stage. For example, if households were selected within primary sampling units (PSUs), then the household base weight would be

$$W_k = \frac{1}{P_{hi} CP_{hik}},$$

where P_{hi} is the probability of selecting PSU i in stratum h , and CP_{hik} is the conditional probability of selecting household k within PSU i of stratum h .

The household selection probability also reflected any duplicate records in the sampling frame or any changes to the subsampling procedures.

Household unknown eligibility adjustment

Before any household-level nonresponse adjustment was applied, an adjustment for unknown eligibility was performed if the eligibility status of some households could not be determined. In this step, a portion of the weights of the households with unknown eligibility status (i.e., whether they contained a person age 16 to 65) was distributed to ineligible cases. An adjustment factor was computed as the proportion eligible among those with known eligibility status to down-weight the cases with unknown eligibility status (accounting for an estimated proportion that was ineligible). The down-weighted unknown eligibility cases were then treated as eligible nonrespondents. This adjustment was done within weighting cells defined for the unknown eligibility adjustment (see Table 15-3).

Household nonresponse adjustment

For the screener nonresponse adjustment, the nonrespondents were divided into two categories. The first consisted of cases involving nonliteracy-related nonresponse. Examples of this category included refusals and nonresponse due to speech impairment. Nonliteracy-related nonrespondents were likely to be similar to respondents with respect to proficiency scores. The second category was literacy-related nonresponse. Language problem was the only type of literacy-related nonresponse at the screener level. Households with this type of nonresponse were presumed to differ from responding households with respect to proficiency. Therefore, the weighting procedures adjusted the weights of the respondents to represent the nonliteracy-related nonrespondents only. The weights of the LRNRs were not adjusted during the screener-level nonresponse adjustment because their proficiency was expected to differ from that of respondents. The contribution of the screener level literacy-related nonresponse to the total population was accounted for by the literacy-related nonresponse adjustment carried out at the person level involving the assessment LRNRs (see section 15.1.3).

The next step in the weighting process was to adjust the unknown eligibility-adjusted weights to reduce potential bias as a result of nonresponse to the screener. An adjustment was made to distribute the screener unknown eligibility-adjusted weights of the nonliteracy-related nonrespondents to the screener respondents. The nonresponse adjustment was performed within cells that were defined based on pre-selected weighting variables that were found to be related to proficiency and to response propensity (see Table 15-3). Within each adjustment cell, the household unknown eligibility-adjusted weights of nonrespondents were redistributed over a relatively large pool of cases (approximately 30 or more respondents). Additionally, the amount of variation in the nonresponse adjustment factors was kept to a minimum by limiting the maximum allowable nonresponse adjustment factor, which was a function of the achieved screener response rate.

15.1.3 Person-level weighting adjustments

This section describes the process of creating the person-level weights, including the computation of person base weights; the person unknown eligibility adjustment that applied to registry countries only; the nonresponse adjustment procedure designed to reduce potential nonresponse bias; the calibration of weights to control totals; and the general trimming procedure used to reduce the impact of extreme weights.

Person base weights

For screener countries, the person base weights accounted for both nonresponse to the household screener and differential within-household selection rates. The person base weights were computed as the product of the household nonresponse-adjusted weight and the reciprocal of the within-household person selection probability.

For registry countries, the base weight for each sampled person was computed as the reciprocal of the person selection probability.

Person unknown eligibility adjustment

For registry countries, an adjustment for person unknown eligibility was performed if the eligibility status of some sampled persons could not be determined due to the inability of the survey to locate and interview these selected persons not residing at the address listed in the registry (see section 16.2.2 for a discussion on inaccessible sampled persons). In the person unknown eligibility adjustment, a portion of the person base weights of the sampled persons with unknown eligibility status was distributed to the ineligible cases. An adjustment factor was computed as the proportion eligible among those with known eligibility status to down-weight the cases with unknown eligibility status (accounting for an estimated proportion that was ineligible). The down-weighted unknown eligibility cases were then treated as eligible nonrespondents in the nonresponse adjustment.

Person nonliteracy-related nonresponse adjustment

For the nonresponse adjustment, the nonrespondents were divided into two categories. The first category consisted of nonliteracy-related nonrespondents (e.g., refusals and inaccessibles with known eligibility) and sampled persons with a disability (e.g., hearing impairment and physical disability). They were likely to be similar to respondents with respect to proficiency scores. The second category was literacy-related nonresponse (LRNR). Types of literacy-related nonresponse include language problem, reading and writing difficulty, and learning-mental disability. Sampled persons with this type of nonresponse were presumed to differ from respondents with respect to proficiency. Therefore, LRNRs received a different treatment than nonliteracy-related nonrespondents.

As mentioned earlier, for screener countries, an adjustment was made to distribute the person base weights of the nonliteracy-related nonrespondents and sampled persons with a disability to the respondents' weights.

For registry countries, excluded inaccessible sampled persons were treated as nonliteracy-related nonrespondents in weighting. An adjustment was made to distribute the person unknown eligibility-adjusted weights of the nonliteracy-related nonrespondents, sampled persons with a disability, and down-weighted unknown eligibility cases to respondents.

The nonresponse adjustment was performed within cells that were defined based on pre-selected weighting variables that were found to be related to proficiency and to response propensity (see Table 15-3). Within each adjustment cell, the person unknown eligibility-adjusted weights of nonrespondents were redistributed over a relatively large pool of cases (approximately 30 or more respondents). Additionally, the amount of variation in the nonresponse adjustment factors

was kept to a minimum by limiting the maximum allowable nonresponse adjustment factor, which depended on the achieved BQ response rate.

Person literacy-related nonresponse adjustment

For screener countries, the weights of the BQ and assessment LRNRs were adjusted to account for the screener LRNRs. This adjustment was necessary primarily to allow both the BQ and assessment LRNRs to represent the screener LRNRs in the calibration procedure. This adjustment assumed that the LRNRs to the screener, BQ and assessment were similar in proficiency.

For registry countries, the weights of the BQ LRNRs with age and gender collected and assessment LRNRs were adjusted to account for the weights of the BQ LRNRs without age and gender collected.

Involving the assessment LRNRs in the literacy-related nonresponse adjustment offered several advantages. This approach (1) reduced the mean square error in the resulting estimates, (2) provided stability in the weight adjustment and reduced the variations in the weights and in the estimates, (3) reduced bias under the assumption that the assessment LRNRs were more similar to the BQ LRNRs than the BQ nonliteracy-related nonrespondents, and 4) addressed the issue that sampled persons may or may not have completed the BQ because of an arbitrary reason (e.g., unavailable bilingual interviewer or interpreter).

Calibration

To address undercoverage bias, to reduce the mean square error of estimates and to create consistency with statistics from other studies, the next weighting step was to adjust the survey weights to match population control totals. At minimum, weights were benchmarked to control totals for age and gender. Respondents who completed the BQ and BQ LRNRs received a final weight and were included in calibration. If the Consortium performed the weighting adjustments, one iteration of calibration, trimming (if necessary) and recalibration was performed following the nonresponse adjustments. Not all countries that performed their own weighting included the initial calibration prior to trimming.

Three main calibration techniques employed by countries are post-stratification, raking and generalized regression estimators (GREG). Post-stratification adjusts survey weights of respondents so that the weighted sample distribution is the same as some known population distribution (i.e., the sums of the adjusted weights of the respondents are equal to known population totals for certain subgroups of the population). The raking procedure uses an iterative procedure to adjust the survey estimates to the known marginal totals of several categorical variables. The GREG estimator is a model-assisted approach that can be used to adjust weights to exploit explicitly the relationship between a survey variable and auxiliary variables.

Trimming the outliers

Even a carefully designed sample could not fully prevent the need for reducing extreme weights. Sample designs that included the selection of dwelling units had more variability in the weights compared to directly sampling persons from registries because of unequal household sizes. The use of nonresponse and calibration adjustments also introduced variations in sampling weights.

Weight trimming introduced some bias into the sampling weights. However, the trimming adjustment in most cases reduced the sampling error component of the overall mean square error more than it increased the bias as the adjustment was applied to only a relatively small number of weights (Lee, 1995).

The person weights were trimmed as necessary after the first calibration. Using a design-based procedure, cells for trimming were formed from groups that were expected to be approximately self-weighting. In each cell, weights above a cutoff value were trimmed down to the designated cutoff. To define the trimming cut point, the Consortium examined the coefficient of variation (CV) based on the weights after raking (the cut point was calculated separately by domain in case oversampling was used for some domains). The Consortium trimmed the weights that were over $3.5 \times \sqrt{1 + CV^2}$ times the median raked weight (within each trimming cell, if sampling rates varied by sampling domains). In a few instances, a review of the distribution of the raked weights revealed that a different cut point was more appropriate. Some countries that performed their own weighting used different criteria for trimming. During trimming, the trimming factor was applied to each replicate weight. After trimming, the weights were recalibrated back to the control totals.

15.1.4 Weighting quality control checks

Quality control (QC) checks were performed for both the full sample and replicate weights after each adjustment in the weighting procedure to ensure proper implementation. The Consortium developed a battery of QC checks to review the weighting process for adherence to the weighting standards and guidelines and to check weight calculations for reasonableness and accuracy. Performing the weighting QC checks was essential for verifying that the final weights produced for estimation are appropriate (see section 16.1). The PIAAC schedule required the weighting QC checks to be conducted prior to the development of proficiency scores. Further checks were conducted after derivation of the proficiency scores if analyses showed any need for re-verification/correction of the weights.

15.1.5 Summary of country-specific weighting implementation

This section presents the weighting steps performed by countries, variables selected by countries for weighting adjustments and country-specific deviations from the weighting standards. All participating countries in PIAAC were responsible for selecting weighting variables and preparing files for weighting. The Consortium was responsible for deriving sampling weights for the Main Study for all countries. Countries that opted to compute their own weights were required to follow the standards and guidelines in Chapter 14 of the PIAAC Technical Standards and Guidelines and the PIAAC Weighting and Variance Estimation Plan. The weighting procedures described in the standards ensured that the estimates represent each country's target population and reduce the potential for bias due to nonresponse.

Weighting steps performed by countries

Table 15-2 indicates each participating country's weighting responsibility, sample design, weighting steps performed, and calibration method. Any deviations from the weighting standards and special weighting adjustments are noted in Table 15-5.

Table 15-2: Weighting steps, by country

Country	Weighting Respon-sibility	Design	Screener			Background Questionnaire					
			Base Weight	Unknown Eligibility Adjustment	Nonresponse Adjustment	Base Weight	Unknown Eligibility Adjustment ¹	Nonresponse Adjustment (nonliteracy-related)	Nonresponse Adjustment (literacy-related) ²	Trimming ³	Calibration
Australia	Country	Screener	Y	N	N	Y		Y	Y	N	GREG
Austria	Westat	Registry				Y	Y	Y	Y	Y	Raking
Canada	Country	Screener	Y	Y	Y	Y		Y	Y	Y	Raking
Cyprus ³	Westat	Screener	Y	Y	Y	Y		Y	Y	Y	Raking
Czech Republic	Westat	Screener	Y	Y	Y	Y		Y	Y	Y	Raking
Denmark	Country	Registry				Y	NA	Y	Y	N	GREG
England (UK)	Westat	Screener	Y	Y	Y	Y		Y	Y ⁴	Y	Raking
Estonia	Westat	Registry				Y	Y	Y	NA	Y	Raking
Finland	Country	Registry				Y	Y	Y	N	Y	GREG
Flanders (Belgium)	Westat	Registry				Y	NA	Y	NA	Y	Raking
France	Westat	Registry				Y	Y	Y	NA	N	Raking
Germany	Westat	Registry				Y	Y	Y	Y	Y	PS
Ireland	Westat	Screener	Y	Y	Y	Y		Y	Y	Y	Raking
Italy	Country	Screener	Y	Y	Y	Y	Y	Y	Y	Y	Raking
Japan	Country	Registry				Y	Y	Y	NA	Y	GREG
Korea	Westat	Screener	Y	Y	Y	Y		Y	Y	Y	Raking
Netherlands	Country	Registry				Y	Y	Y	Y	N	GREG
N. Ireland (UK)	Westat	Screener	Y	Y	Y	Y		Y	Y ⁴	Y	Raking
Norway	Country	Registry				Y	Y	Y	Y	Y	Raking
Poland	Westat	Registry				Y	Y	Y	N	Y	Raking
Russian Federation ⁴	Westat	Screener	Y	Y	Y	Y		NA	NA	Y	Raking
Slovak Republic	Westat	Registry				Y	NA	Y	Y	Y	Raking
Spain	Country	Registry				Y	Y	Y	NA	Y	GREG
Sweden	Country	Registry				Y	Y	N	Y	N	GREG
United States	Country	Screener	Y	Y	Y	Y		Y	Y	Y	Raking

]: not applicable, Y: weighting step performed, N: weighting step not performed, NA: weighting step not needed, PS: post-stratification

¹* NA: There were no cases with unknown eligibility status (i.e., DISP_CIBQ=24 and EXCFLG=2).

² NA: There were no LRNRs with age and gender not collected (i.e., DISP_CIBQ = 7, 8, or 9 and QCFLAG_LR = 2) or no LRNRs at the screener level (DISP_SCR=7).

³ A value of “Y” indicates that the weighting process included a step to evaluate whether there were any extreme weights and trim if necessary. It does not indicate the outcome of the trimming (i.e., whether any weights were trimmed).

⁴ In addition to the standard literacy-related nonresponse adjustment, LRNRs with age and gender successfully collected represented those with age or gender not successfully collected.

³ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

⁴ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Weighting variables selected by countries

After data collection and data editing, countries were to conduct an analysis to select variables for weighting adjustments that would be most effective in reducing nonresponse bias. At minimum, this analysis was to involve a classification tree or logistic regression to evaluate the relationship of response status to potential weighting variables.

The list of weighting variables selected by each country is given in Table 15-3. Of the countries that provided information, all used age and gender in calibration, as required in the PIAAC Technical Standards and Guidelines, and region was also used in all countries in either calibration or nonresponse adjustment. In addition, the majority of countries included in their weighting adjustments at least one variable related to education, employment status or nationality, which have been shown to be correlated with proficiency.

Benchmark control totals used by countries

Control totals used in the benchmarking process were required to have the same definition and coverage of the target population as PIAAC (noninstitutionalized adults who are between age 16 and 65, including citizens and noncitizens). If not, the counts from the external sources needed to be adjusted to make these comparable to the survey estimates. All variables selected for benchmarking must have reliable control totals available. The quality of data from external sources must have exceeded the quality of data from PIAAC (e.g., the standard errors, or more generally, the mean square error of the external estimates needed to be smaller than those of the nonbenchmark estimates from the survey). Table 15-4 presents the control total variables used in calibration for each country, including its source and exclusions from the target population.

Table 15-3: Weighting variables, by country

Country	Screeners Nonresponse Adjustment	Unknown Eligibility Adjustment	BQ Nonresponse Adjustment (nonliteracy- related)	BQ Nonresponse Adjustment (literacy- related)	Calibration
Australia	NA	NA	1 Cell	1 Cell	Highest educational attainment by state, labor force status by state by sex, labor force status by age group, state by part of state by sex by age group
Austria		Age by citizenship by education by urbanization (8)	Age by citizenship by education by urbanization (8)	Age by citizenship by education by urbanization (8)	Region by age (90), region by citizenship (18), region by level of urbanization by sex (48), sex by age by education (40)

Table 15-3 (cont.): Weighting variables, by country

Country	 Screener Nonresponse Adjustment	 Unknown Eligibility Adjustment	 BQ Nonresponse Adjustment (nonliteracy-related)	 BQ Nonresponse Adjustment (literacy-related)	 Calibration
Canada	2011 Canadian Census short form (2A) questions and census paradata, 2006 census long form (2B) data at geographically aggregated level	?	The variables used for the screener NR adjustment were used. In addition, age and gender of the selected persons was used (333)	Delineation between general population and special subpopulations sample by province (30)	Age group and gender by province (130), educational attainment by province (52), immigration status and gender by province (21), aboriginal status and gender by province (24), census metropolitan area by province (26), linguistic minority status and gender by province (17)
Cyprus ⁶	District by locale (7)	District by locale (9)	District, locale, age, education, gender (34)	District by locale (9)	Age by district (25), age by gender (10), age by education (15), gender by district (10), gender by education (6), language (2)

⁶ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 15-3 (cont.): Weighting variables, by country

Country	 Screener Nonresponse Adjustment	 Unknown Eligibility Adjustment	 BQ Nonresponse Adjustment (nonliteracy-related)	 BQ Nonresponse Adjustment (literacy-related)	 Calibration
Czech Republic	Region (8), municipality type (3), gender ratio quartiles (4), age ratio quartiles (4), employment status percentage quartiles (4), entrepreneurs percentage quartiles (4), education quartiles (4)	Region (8), municipality type (3), gender ratio quartiles (4), age ratio quartiles (4), employment status percentage quartiles (4), entrepreneurs percentage quartiles (4), education quartiles (4)	Municipality type (3), region (8), gender (2), age group (5), employment status percentage quartiles (4), entrepreneurs percentage quartiles (4), education quartiles (4)	1 Cell	Age by education (15), age by gender (10), education by gender (8), field of study by gender (16), work status by gender (14), region by employment status (24), region by education (32)
Denmark		NA	Income, region, education, type of family, mobility, marital status, socio-economic status, employment, gender (70)	1 Cell	Region (5), age (5), gender (20), immigration (4)

Table 15-3 (cont.): Weighting variables, by country

Country	 Screener Nonresponse Adjustment	 Unknown Eligibility Adjustment	 BQ Nonresponse Adjustment (nonliteracy-related)	 BQ Nonresponse Adjustment (literacy-related)	 Calibration
England (UK)	Region (9), National Statistics 2001 Area Classification (21), index of multiple deprivation split into approximate deciles (10), 2001 census percentage living in social housing (9), 2001 census percentage Black or South Asian (7), 2001 census percentage of households that contain one person (10)	Region (9), National Statistics 2001 Area Classification (21), Index of multiple deprivation split into approximate deciles (10), 2001 census percentage living in social housing (9), 2001 census percentage Black or South Asian (7), 2001 census percentage of households that contain one person (10)	Region (9), national statistics 2001 area classification (21), index of multiple deprivation split into approximate deciles (10), 2001 census percentage living in social housing (9), 2001 census percentage Black or South Asian (7), 2001 census % of households that contain one person (10)	Region (9), national statistics 2001 area classification (21), index of multiple deprivation split into approximate deciles (10), 2001 census percentage living in social housing (9), 2001 census percentage Black or South Asian (7), 2001 census percentage of households that contain one person (10)	Gender by age (20), region (9), age by qualifications (17), gender by age by economic status (35)
Estonia		Age (2), gender (5), mother tongue (2), urbanization (3), county (15), percent of high education (4), percent of unemployment (4)	Age (2), gender (5), mother tongue (2), urbanization (3), county (15), percent of high education (4), percent of unemployment (4)	Age (2), gender (5), mother tongue (2), urbanization (3), county (15), percent of high education (4), percent of unemployment (4)	Gender by age (10), county (15), urbanization (3)
Finland		?	Gender (2), age (5), education (4), native language (3), region (5), urban/rural (3), family status (5)	Gender (2), age (5), education (4), native language (3), region (5), urban/rural (3), family status (5)	Gender (2), age (5), education (4), native language (3), region (5), urban/rural (3), family status (5)

Table 15-3 (cont.): Weighting variables, by country

Country	Screeners Nonresponse Adjustment	Unknown Eligibility Adjustment	BQ Nonresponse Adjustment (nonliteracy- related)	BQ Nonresponse Adjustment (literacy- related)	Calibration
Flanders (Belgium)			Age (5), gender (2), province (5)		Age by work status (10), gender by work status (4)
France		Gender (2), age (5), region (3), income (5)	Gender (2), age (5), region (3), income (5)	Gender (2), age (5), region (3), income (5)	Age by gender (10), region (3), education (3), country of birth (2), employment status (3)
Germany ¹		Age, nationality, degree of urbanization	Age, nationality, degree of urbanization	1 Cell	Age, gender, region, education
Ireland	Percentage non- English language spoken at home (2), percentage unemployment (2), percentage with lower secondary-level education or below (2), owner occupied (2), regions (3)	Percentage non- English language spoken at home (2), percentage unemployment (2), percentage with lower secondary-level education or below (2), owner occupied (2), regions (3)	Gender (2), age (5), education (screener) (13)	Gender (2), age (5), education (screener) (13)	Region by age (40), region by gender (16), age by education (20), gender by education (8)

Table 15-3 (cont.): Weighting variables, by country

Country	 Screener Nonresponse Adjustment	 Unknown Eligibility Adjustment	 BQ Nonresponse Adjustment (nonliteracy-related)	 BQ Nonresponse Adjustment (literacy-related)	 Calibration
Italy	Deciles of logit from model involving: Number of eligible persons in family, gender, age, municipality MOS, self-representing PSU indicator, region (10)	Quintiles of logit from model involving: Number of eligible persons in family, gender, age, municipality MOS, self-representing PSU indicator, region (5)	Number of eligible persons in family, gender, age, municipality MOS, self-representing PSU indicator, region (30)	1 Cell	Region by age (25), region by gender (10), region by education (15), region by employment (10)
Japan		Age (5), gender (2)	Age, gender, city size, region, type of building, area-level percentage: graduate from college, population density, household floor space, percentage of people employed in tertiary industry, number of persons per household, proportion of temporary workers to regular employees (20)	Age (5), gender (2)	Age (5), gender (2), education (6), employment status (3), geographic area (10)
Korea	Korea	Region (16), household type (3)	Region (16), household type (3)	Region (16), age (6), gender (2)	1 Cell

Table 15-3 (cont.): Weighting variables, by country

Country	Screeners Nonresponse Adjustment	Unknown Eligibility Adjustment	BQ Nonresponse Adjustment (nonliteracy- related)	BQ Nonresponse Adjustment (literacy- related)	Calibration
Netherlands		Origin (3), household composition (5), social status (3), social status (3)	Origin (3), household composition (5), social status (3), social status (3)	Origin (3), household composition (5), social status (3)	Gender by age (10), origin by generation (5), group of provinces by degree of urbanization (18), household type (5), social status by income (25), term of registration in population registry (2), percentage of high level education by percentage of low level education (18)
Norway		?	Education, occupation, age group, industry and “special field”	?	Gender by age (10)

Table 15-3 (cont.): Weighting variables, by country

Country	 Screener Nonresponse Adjustment	 Unknown Eligibility Adjustment	 BQ Nonresponse Adjustment (nonliteracy-related)	 BQ Nonresponse Adjustment (literacy-related)	 Calibration
Northern Ireland (UK)	Region (5), National Statistics 2001 Area Classification (20), 2001 census percentage living in social housing (9), index of multiple deprivation split into approximate deciles (10)	Region (5), National Statistics 2001 Area Classification (20), 2001 census percentage living in social housing (9), index of multiple deprivation split into approximate deciles (10)	Region (5), National Statistics 2001 Area Classification (20), 2001 census percentage living in social housing (9), index of multiple deprivation split into approximate deciles (10)	Region (5), National Statistics 2001 Area Classification (20), 2001 census percentage living in social housing (9), index of multiple deprivation split into approximate deciles (10)	Gender by age (20), region (5), age by qualifications (17), gender by age by economic status (35)
Poland		Income (4), age (5), population (9), region (16), number of cities per county (11), level of unemployment (5), proportion of middle-school students (4), computerization (4)	Income (4), age (5), population (9), region (16), number of cities per county (11), level of unemployment (5), proportion of middle-school students (4), computerization (4)	Income (4), age (5), population (9), region (16), number of cities per county (11), level of unemployment (5), proportion of middle-school students (4), computerization (4)	Gender by age (10), gender by region (32)

Table 15-3 (cont.): Weighting variables, by country

Country	Screeners Nonresponse Adjustment	Unknown Eligibility Adjustment	BQ Nonresponse Adjustment (nonliteracy-related)	BQ Nonresponse Adjustment (literacy-related)	Calibration
Russian Federation ⁷	Macro-region (8), type of settlement (3), type of district (3), education rate (3), unemployment rate (3)	Macro-region (8), type of settlement (3), type of district (3), education rate (3), unemployment rate (3)	NA	NA	Gender by age (20), education rate (3), macro-region (8)
Slovak Republic			Size of municipality (9), urban/rural (2), region (8), age by gender (10)	1 Cell	Size of municipality (9), urban/rural (2), region (8), age by gender (10)
Spain		Age (5), gender (2), nationality (2)	Age (5), gender (2), nationality (2), urbanicity (3), education (3), unemployment rate (4)	BQ LRNRs (1)	Gender (2), age (5), region (18), nationality (2), education (3)
Sweden		NA	NA	?	Education by sex by age (30), education by region (24), education by employment (9), education by income (12), education by country of birth (6)

⁷ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 15-3 (cont.): Weighting variables, by country

Country	 Screener Nonresponse Adjustment	Unknown Eligibility Adjustment	BQ Nonresponse Adjustment (nonliteracy- related)	BQ Nonresponse Adjustment (literacy- related)	Calibration
United States	Metropolitan Statistical Area, region, categorized percent of: Housing units occupied by owner, Hispanic or non-Hispanic Black, Hispanic, population age 18-64 unemployed, population below 150% of poverty, foreign born, household linguistically isolated, population age 25+ with high school education, population age 25+ with some college education, categorized household size (26)	Metropolitan Statistical Area, region, categorized percent of: Housing units occupied by owner, Hispanic or non-Hispanic Black, Hispanic, population age 18-64 unemployed, population below 150% of poverty, foreign born, household linguistically isolated, population age 25+ with high school education, population age 25+ with some college education, categorized household size (26)	Metropolitan Statistical Area, region, categorized percent of: Housing units occupied by owner, population age 25+ with at least high school education, Hispanic or non-Hispanic Black, Hispanic, population age 18-64 unemployed, foreign born, household linguistically isolated, population age 18-64 employed, population age 25+ with some college education, categorized household size, best age category (after imputation), indicator for children under age 16 in household, best gender, best race/ethnicity (after imputation) (26)	1 Cell	Educational attainment by race/ethnicity (12), education attainment by age (20), education attainment by gender (8), race/ethnicity by age (9), race/ethnicity by gender (6), country of birth by age (10), country of birth by region (8)

□ not applicable, NA: weighting step not performed, ?: unknown/received no information from country

¹ The number of categories is not provided for confidentiality reasons.

NOTE: Numbers in parentheses indicate the number of categories.

Table 15-4: Benchmark control totals, by country

Country	Population Total	Source	Year	Exclusion From Control Totals
Australia ¹	16,704,354 (age 15-74)	Estimated resident population, projected from Census	2006	None
		Monthly Population Survey (MPS)	2011-2012	Members of the permanent defense forces, certain diplomatic personnel of overseas governments customarily excluded from census and estimated population counts, overseas residents in Australia, and members of non-Australian defense forces (and their dependents) stationed in Australia
		Survey of Education and Work (SEW)	2011	Ages 65-74, special dwelling type institutionalized persons, special dwelling type boarding school pupils, persons permanently unable to work, and persons living in collection districts that contain a discrete indigenous community in very remote areas
Austria	5,647,341	Population registry and Labor Force Survey	2011	Undocumented immigrants
Canada	23,381,067	Demographic projections of the Canadian population for April 2012 based on 2006 Census data	2012	Indian reserves in the provinces, institutions and non-institutional collective dwellings
Cyprus ⁷	592,296	Census	2011	None
Czech Republic	7,395,111	Census	2011	Undocumented immigrants
Denmark	3,629,087	Registry	2011	Undocumented immigrants
England (UK)	34,257,191	Simple mean values for population estimates produced for each quarter in the calendar year 2011	2011	None
Estonia	896,163	Official Demographic Statistics	2012	Undocumented immigrants

⁷ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 15-4 (cont.): Benchmark control totals, by country

Country	Population Total	Source	Year	Exclusion From Control Totals
Finland	3,496,909	Population database, education register for education level	2011	None
Flanders (Belgium)	4,138,042	Labor Force Survey	2010	None
France	4,0793,515	Labor Force Survey	2012	None
Germany	53,657,540	Microcensus	2010	Undocumented immigrants
Ireland	2,994,368	Census	2011	None
Italy	39,369,830	Italian Multipurpose Survey	2010	None
Japan	81,059,238	Census	2010	None
Korea	34,602,008	Census	2010	Undocumented immigrants, residents of small islands
Netherlands	11,160,541	Registry	2011, 2011-2012	Non-registered population
Northern Ireland (UK)	1,165,218	March 2010 population estimates	2010	None
Norway	3,282,755	Registry	2011	Undocumented immigrants
Poland	26,741,987	Registry	2011	Undocumented immigrants and foreigners staying in Poland fewer than 3 months
Russian Federation ⁸	87,415,088	Census	2010	Moscow region and Moscow city
Slovak Republic	3,870,993	Census	2011	None
Spain	31,091,563	Registry	2012	None
Sweden	6,116,358	Registry	2011	Undocumented immigrants
United States	203,144,374	American Community Survey	2010	None

¹ Control totals were adjusted to meet the PIAAC scope, that is, all persons aged between 15 and 74 years old who do not live in very remote areas, special (i.e., nonprivate) dwellings, or collection districts that contain a discrete indigenous community, and exclude persons that are diplomatic personnel of overseas governments.

⁸ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Deviations from weighting standards or excluded steps

The majority of countries performed each of the weighting steps described in section 15.1.2 and/or 15.1.3. The exceptions are enumerated in Table 15-5.

Table 15-5: Deviations from weighting standards or excluded steps, by country

Country	Weighting Deviations or Excluded Steps
Australia	Australia used person-level nonresponse adjustments and benchmarking to adjust for undercoverage and nonresponse at the household and person level, rather than performing a series of separate adjustments. Australia also applied an explicit trimming step, but if a weight was lower than 50% or higher than 300% of the initial weight after adjustments and benchmarking, benchmark classes were collapsed to reduce the weight fluctuation.
Austria	None
Canada	Canada's sample included several oversamples that were selected sequentially from the 2011 Canadian census or the 2011 National Household Survey databases, meaning that (1) there was an overlap between the frames used to select each sample, and (2) a unit selected for one part of the sample was no longer available for the other parts of the sample. As a result, the sum of weights of the whole sample would overestimate the size of the Canadian population aged between 16 and 65. Canada included an integration step at the end of the weighting process so that the final weights adequately represent the PIAAC population.
Cyprus ⁹	None
Czech Republic	Weights for the Czech Republic main sample and supplemental sample were created separately and then composited at the end of the weighting process. In the supplemental sample, 30-year-olds were treated as 29-year-olds. The main, reserve and supplemental sample were selected in a sequential manner, and the screener base weights for the reserve and supplemental samples reflected conditional probabilities given the household was not selected for the previous sample. Therefore, the base weights for the sample main sample (including reserve) were adjusted downward so that they sum to the total of the base weights of the main sample without reserve. Following compositing, the weights for the combined samples were raked to ensure that the final composited weights agreed with the control totals used when raking the main sample.
Denmark	An unknown eligibility adjustment was not needed because Denmark did not have any inaccessible cases with unknown whereabouts.
England/N. Ireland (UK)	England/N. Ireland (UK) did not collect age and gender for all sampled persons during the screener. Therefore, in addition to the standard literacy-related nonresponse adjustment for screener countries, LRNRs with age and gender successfully collected represented those with age or gender not successfully collected. In addition, the theoretical person base weights (THEOR_PBWT) were derived from imputed values of the number of eligible people in the sampled household (NUM_ELG) for some cases due to a technical problem with the contact data that the interviewers entered.
Estonia	A literacy-related nonresponse adjustment was not needed for Estonia because all LRNRs had age and gender collected.
Finland	None

⁹ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 15-5 (cont.): Deviations from weighting standards or excluded steps, by country

Country	Weighting Deviations or Excluded Steps
Flanders (Belgium)	An unknown eligibility adjustment was not needed because Flanders (Belgium) did not have any inaccessible cases with unknown whereabouts. A literacy-related nonresponse adjustment was not needed for Flanders (Belgium) because all LRNRs had age and gender collected.
Germany	Although the sample was probability based, Germany was unable to calculate exact selection probabilities due to an error in the sample selection algorithm. Therefore, the base weights were calculated using estimated probabilities from a simulation.
Ireland	None
Italy	None
Japan	A literacy-related nonresponse adjustment was not needed for Japan because all LRNRs had age and gender collected.
Korea	None
Netherlands	None
Norway	None
Poland	Poland did not collect age and gender for any of the BQ LRNRs and had very few assessment LRNRs, so the standard literacy-related nonresponse adjustment could not be performed. The BQ LRNRs together with the other BQ NRs were represented by BQ respondents. Poland's data were reweighted to correct for base weights. Poland discovered after weighting that in four cities the sample was not selected with equal probability (base weights adjusted to reflect differential selection probability) and a city was omitted during sample selection (base weights inflated for other cities with similar population to represent the omitted city). This led to more variability in their final weights.
Russian Federation ¹⁰	A literacy-related nonresponse adjustment was not needed for the Russian Federation because there were no literacy-related nonrespondents at any stage of the data collection. Also, BQ nonresponse adjustment was not conducted because the BQ response rate was close to 100%.
Slovak Republic	An unknown eligibility adjustment was not needed because the Slovak Republic did not have any inaccessible cases with unknown whereabouts.
Spain	A literacy-related nonresponse adjustment was not needed for Spain because all LRNRs had age and gender collected.
Sweden	Sweden used benchmarking to adjust for undercoverage and nonresponse rather than performing a series of separate adjustments. To meet the requirements for the appropriate treatment of LRNRs, Sweden inflated the weights of assessment LRNRs to account for BQ LRNRs without age and gender collected. Then the base weights for the respondents were calibrated directly to known population totals (less the total for the LRNRs). Data collected from the survey (e.g., age) were not used in weighting, as all weighting variables were based on the registry data. After calibration, Sweden performed an unknown-eligibility adjustment to adjust for ineligible cases since their population totals included ineligible cases.
United States	None

¹⁰ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

15.2 Variance estimation

Inferences will not be valid unless the corresponding variance estimators appropriately reflect all of the complex features of the PIAAC sample design (e.g., stratification and clustering). The replication approach is used for estimating variances for the international analyses of PIAAC data. Under the replication approach, subsamples (also known as replicates) from the full sample are formed and statistics of the subsamples are used to estimate the variance of the full sample statistic. The replication approach, in conjunction with the multiple imputation approach used to derive the plausible values, captures the variation due to the complex sampling and estimation approaches, including:

- Sample design
- Selection
- Weighting adjustments
- Measurement error through the processing of multiple imputation of plausible values

For a detailed description on replication methods for different sample designs, refer to Appendix D of the WesVar® manual.¹¹

The PIAAC Data Explorer is the primary tool for the analysis of PIAAC data. It has been adapted for handling the following four different replication schemes:

- Delete-one jackknife
- Paired jackknife
- Balanced repeated replication
- Fay's method

The delete-one jackknife is also referred to as delete-a-group jackknife, random groups approach or JK1. The paired jackknife is also referred to as JK2. The JK2 approach, with two variance units per stratum, is appropriate for sample designs where PSUs are stratified or selected with systematic sampling from a sorted list. The balanced repeated replication (BRR) approach is also commonly used when strata are involved, and Fay's method is a variant of the BRR approach.

Replication methods are applied to surveys by dividing the sample into specially designed replicate subsamples that mirror the design of the full sample. To form the replicate subsamples, variance strata and variance units are defined. Each subsample is reweighted to account for the subsampling that occurred. An estimate is then calculated for the full sample and each of the replicate subsamples. The variance of the full sample estimate is computed as the sum of squared deviations between each replicate subsample estimate and the full sample estimate. The general replication formula is

$$Var(\hat{\theta}) = c \sum_i (\hat{\theta}_i - \hat{\theta}_0)^2$$

where

¹¹ http://www.westat.com/Westat/pdf/wesvar/WV_4-3_Manual.pdf

c	=	1,	for the paired jackknife (JK2)
	=	$(g-1)/g$,	for the random groups (delete-one) approach (JK1)
	=	$1 / g$	for the BRR approach
	=	$1/[g(1-k)^2]$	for Fay's method
g	=	number of replicates	
k	=	weighting factor for Fay's method	
$\hat{\theta}_0$	=	full sample estimate	
$\hat{\theta}_i$	=	estimate for replicate i .	

A variety of sample designs were employed across the different countries participating in PIAAC. Replication is adaptable to a wide variety of designs, including simple random sampling, systematic sampling, stratified designs and multistage cluster designs. In general, replication schemes are selected based on the sample design. A random groups approach may do well for a simple random sample while a paired jackknife mechanism is not meant for an SRS, but could be adapted. The paired jackknife would work very well for a one-PSU per stratum design, while a random groups design is not appropriate. Some efficiency is gained by selecting the most appropriate approach for the sample design.

15.2.1 Creation of replicate weights

Participating countries followed the PIAAC Technical Standards and Guidelines in providing the data necessary for creating replicate weights. All participating countries in PIAAC were responsible for defining variance strata and variance units. The specification of variance strata and variance units must conform to the design assumptions of a replication method and should be determined by the type of sampling design that was used to collect the data (e.g., whether or not stratification was used and how many PSUs were in each stratum). In addition, in some cases the sampling strata and PSUs had to be grouped to reduce the number of replicates to fit the sample design into a replication design that followed the PIAAC standards.

Once the variance strata and variance units were assigned, the Consortium/countries followed detailed guidelines on how to form and create the replicate weights. First, replicate base weights were created. For screener countries, the household base weights for the household were replicated. For registry countries, the person base weights were replicated. Subsequently, all weight adjustments that were conducted for the full sample were conducted on each replicate weight to capture the variation created, or reduced, by the weight adjustments.

15.2.2 Summary of country-specific variance estimation implementation

Table 15-6 presents the replication approach employed by each country. The choice of the replication method was guided by the particular sample design used in each country. For instance, JK1 is appropriate for a design that uses a registry without stratification or sorting. If strata were used and there were two primary sampling units (PSUs) per stratum, the appropriate

replication method would be JK2, BRR or Fay’s method. If there were many PSUs sampled from a small number of strata, then JK2, BRR or Fay’s method could still have been used to reflect the sampling variation by creating pseudo-strata within the existing strata. The allowed number of replicates ranged from a minimum of 15 to a maximum of 80 replicate weights.

Table 15-6: Replication approach, by country

Country	First Stage Sample Design		Replication Method	Number of Replicates
	Stratification	Number of Sampled Units Per Stratum (for non-certainties)		
Australia	Yes	Not reported	JK1	60
Austria	Sorting only	NA	JK1	80
Canada	Yes	More than 2	JK1	80
Cyprus ¹²	Yes	More than 2	JK2	80
Czech Republic	Yes	More than 2	JK2	80
Denmark ¹	Yes	More than 2	JK1	80
England (UK)	Yes	More than 2	JK2	80
Estonia	Yes	More than 2	JK2	80
Finland	Yes	More than 2	JK2	80
Flanders (Belgium)	Yes	More than 2	JK2	80
France	Yes	More than 2	JK2 ²	80
Germany	Yes ³	0, 1, or 2	JK1	80
Ireland	Yes	More than 2	JK2	80
Italy	Yes	2	JK2	80
Japan	Yes	More than 2	JK2	80
Korea	Yes	More than 2	JK2	80
Netherlands	Sorting only	NA	JK2	80
Northern Ireland (UK)	Sorting only	NA	JK2	80
Norway	Yes	More than 2	JK2	80
Poland	Yes	More than 2	JK2	80
Russian Federation ¹³	Yes	1, 2, 3, or 4	JK2	12 ⁴
Slovak Republic	Yes	More than 2	JK2	80
Spain	Yes	More than 2	JK2	80
Sweden	Yes	More than 2	JK2	80
United States	Yes	1	JK2	45

NA: not applicable; JK1: delete-one jackknife; JK2: paired jackknife.

¹ Denmark discovered an error in the calibration step after weighting had been completed (i.e., some population counts for the replicate calibration program were incorrect). The difference between the erroneous and the correct calibrated weights was less than 0.017 because the procedure calibrated to the correct population total. Because the impact on variances appeared to be small, no re-calibration was warranted.

² France’s replicate weights were created using Fay’s method. However, the variance computation can use the JK2 formula.

³ Germany had a highly stratified design, with more strata than sampled PSUs.

⁴ Due to the small number of PSUs selected, only 12 replicates could be formed for Russian Federation (11 from 22 noncertainty PSUs and 1 from 1 certainty PSU).

¹² Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

¹³ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

15.2.3 Accounting for imputation error variance component

For estimation using plausible values (PVs), calculations must account for both the sampling error component and the variance due to imputation of proficiency scores. The estimator of the population mean is the average of the M PV means,

$$\hat{Y}^* = \sum_{m=1}^M \hat{Y}_m / M.$$

The variance of the estimated mean \hat{Y}^* is computed using formulas specific to PVs as follows:

$$v(\hat{Y}^*) = U^* + B \left(1 + \frac{1}{M}\right)$$

where, the “within” variance component is computed as the average of the sampling variance for each of the M plausible values, computed as,

$$U^* = \left(\sum_{m=1}^M U_m\right) / M,$$

where the sampling variance of the estimated mean \hat{Y}_m for plausible value m is U_m , and

where, the “between” component is calculated as

$$B = \left[\sum_{m=1}^M \left(\hat{Y}_m - \hat{Y}^*\right)^2 \right] / (M - 1)$$

where, the mean of each of the M PVs $y_{l1}, y_{l2}, \dots, y_{lm}$ for sample unit l is computed as

$$\hat{Y}_m = \sum_{l \in s} w_l y_{lm} / \sum_{l \in s} w_l ; m = 1, \dots, M,$$

where s denotes the set of sample units.

The standard error is computed as the square root of the total variance, $\sqrt{v(\hat{Y}^*)}$.

15.3 Recommendations for future cycles

Based on the Field Test and Main Study experience of PIAAC Round 1, the Consortium is proposing a series of recommendations for future cycles of PIAAC.

1. Countries should review the Weighting and Variance Estimation document during data collection and develop the programs needed for the completion of the Sample Design International File (SDIF).
2. More extensive quality checks should be conducted before countries submit the SDIF. The Consortium can provide such checks so they can be implemented by the country and/or incorporated into the Data Management Expert software.
3. Due to the complexities surrounding the assignment of the variance strata and variance units, for which the replicate weights are created, it is recommended that the Consortium conduct the assignment.

4. The Consortium will compute sample weights for all countries to ensure standardization unless a country has a reasonable justification (e.g., confidentiality issues) for weighting its own data.
5. Countries should review the set of variables used in weighting by other countries (Table 15-3) to see if any variables can be added to the weighting process for their country.
6. The same programs used for doing weight adjustments for the full sample weight must be used (or looped through) for each of the replicate weights. If the replicate weights are out of alignment with the full sample weights, it causes significant increase to the variances. This concern is dampened due to the recommendation that the Consortium conduct the weighting.
7. Countries need to ensure that the categories of the calibration variables, to be provided in the SDIF, are exactly the same (in terms of values and meaning) as given in the control totals.
8. Countries should conduct a comparison of control totals for two difference sources, explain the difference, and determine what is needed to be done for the control totals to have the same representation as the PIAAC target population.

References

- Deming, W. E., & Stephan, F. F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, pp. 427-444.
- Lee, H. (1995). Outliers in business surveys. In B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge & P. Kott (Eds.), *Business Survey Methods* (pp. 503-526). New York, NY: John Wiley & Sons.
- Sarndal, C. E., Swenson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, NY: Springer-Verlag.

Chapter 16: Indicators of the Quality of the Sample Data

Leyla Mohadjer, Tom Krenzke and Wendy Van de Kerckhove, Westat

The sampling and weighting procedures described in Chapters 14 and 15 were undertaken with the goal of minimizing total survey error and producing samples that are representative of the target population. This chapter begins with a discussion of the quality assurance and quality control procedures that were implemented to ensure the sampling and weighting standards were met. The remaining sections report key quality indicators for each country. Section 16.2 provides coverage rates and response rates, section 16.3 describes the results of nonresponse bias analyses, and section 16.4 gives sample sizes and design effects.

16.1 Quality assurance and quality control procedures

Quality assurance (QA) and quality control (QC) procedures were put in place to ensure high-quality data that are comparable between countries. Section 16.1.1 describes the sampling-related QA process used by the Consortium to help achieve this goal. Section 16.1.2 describes the QC procedures required of countries to check that the quality goals related to sampling were met. Country compliance with the sampling, weighting, and nonresponse bias analysis QC procedures is addressed in sections 14.7, 15.1.5, and 16.3, respectively.

16.1.1 Quality assurance activities

The QA process for sampling activities involved the development of standards and guidelines, production of sampling documents, creation of sampling and weighting activity toolkits, and communication with countries. This section provides a summary description of each activity.

Technical standards and guidelines for sampling and weighting

For Chapters 4 and 14 of the PIAAC Technical Standards and Guidelines, the Consortium produced standards, guidelines and recommendations for each of the following:

- **Target population:** To ensure that the target population for PIAAC is clearly defined in each country and is consistent across countries
- **Sampling frame:** To ensure that the sampling frame(s) is of high quality, provides acceptable coverage of the target population, and meets the requirements for sampling, location of selected population members, and estimation
- **Sample size:** To establish minimum sample size requirements for each country in order to meet the analysis goals of PIAAC
- **Sample design:** To specify the PIAAC sample design that will produce a probability-based sample, representative of the target population, in each participating country

- **Country-specific supplemental samples:** To describe potential country-specific supplemental sampling options and their implications for sample size
- **Sample selection:** To specify procedures for selecting a probability-based sample from the PIAAC target population following the sample design of PIAAC
- **Indicators of survey quality – noncoverage bias, nonresponse bias, and response rates:** To establish indicators to measure the quality of PIAAC survey data with respect to representation of the target population, and to provide standard procedures for measuring these indicators
- **Respondent incentives:** To increase response rates by offering sampled adults some incentive for participating in PIAAC and for attempting the assessment
- **Sample monitoring:** To monitor the sample during data collection, allowing timely reaction to any developing shortfalls or other potential for bias in the outcome sample
- **Weighting:** To provide a standard weighting approach and to facilitate the production of point estimates for the target population and their associated sampling error estimates

Sampling, weighting and nonresponse bias analysis (NRBA) documents

The Consortium created sampling, weighting and NRBA documents to provide further details on the quality standards in Chapters 4 and 14 of the PIAAC Technical Standards and Guidelines. The PIAAC Sampling Plan for the Field Test and PIAAC Sampling Plan (Main Survey) Part I gave an overview of the PIAAC sample design and a description of the information that countries should include in their sampling plan forms (described below). The PIAAC Weighting and Variance Estimation Plan described the weighting process, including the weighting steps, treatment of different disposition codes, calculation of weighting adjustment factors, assignment of variance strata and variance units, and creation of replicate weights. The paper entitled “PIAAC: Reducing Nonresponse Bias and Preliminary Nonresponse Bias Analysis” described the goals for identifying and reducing nonresponse bias before, during and after data collection. It also included requirements for the NRBA and examples of analyses conducted for past adult literacy surveys.

Sampling and Weighting Activity Toolkits

The Sampling and Weighting Activity Toolkits are a set of Consortium-developed programs and worksheets to aid countries in various sampling- and weighting-related activities. The toolkits were optional to countries but served to provide assistance to countries that needed it and helped ensure consistent and high quality results.

Types of toolkits included are as follows:

- Design effects (DEFF): Excel spreadsheets to compute DEFF due to clustering as well as DEFF due to differential sampling rates
- Within-household selection: Test input files for the algorithm to select one or two persons in a household
- Response rates: Excel spreadsheets to calculate actual and projected response rates for each data collection stage
- Variable selection: Programs, documentation, examples and test files for the selection of weighting variables

- Range of bias: Excel spreadsheet to evaluate the potential for nonresponse bias based on assumptions on how different nonrespondents are from respondents within the weighting classes

Sampling workshops and other communications

Communication with countries is an essential part of the QA process. To this end, the Consortium conducted a sampling workshop at the Barcelona (Spain) NPM meetings in March 2009. The workshop covered information on sample design, sampling plan forms, Field Test sampling requirements and sample sizes, Field Test quality control (QC) forms for sample selection and sample monitoring, and within-household selection. A second sampling workshop was held at the Princeton (NJ) NPM meeting in December 2010 that focused on lessons learned from the Field Test and preparing countries for the Main Study tasks of sample design and selection, weighting and variance estimation, and NRBA.

In February and March 2012, the Consortium held Web meetings to introduce the weighting QC forms (described below) and answer any weighting questions from countries. The sessions were offered at five different dates/times to accommodate country schedules. The Consortium also communicated with countries through presentations on sampling and survey operations requirements at NPM meetings and provided feedback through in-person consultation sessions (at NPM meetings) or through emails as needed.

16.1.2 Quality control activities

Sampling QC checks gathered information necessary to monitor the countries' sampling activities and facilitated a series of validity checks conducted by the Consortium. They were implemented through a series of electronic forms and data files for the Field Test and Main Study. The QC process started with the Consortium reviewing the materials and responding back to the country with suggestions for changes or recommendations for improvements. Each QC form or file had a submission schedule to ensure countries met the timeline for various project activities. Real-time monitoring of all aspects of sampling was critical in allowing the Consortium to uncover problems with sampling activities and for the countries to incorporate changes if necessary.

This section provides a summary description of each QC activity.

Sampling, weighting and NRBA plans

To reduce burden, the Consortium created a series of Sampling Plan Forms that contained all the information needed to meet the requirements listed in Chapter 4 (sample design and selection) and Chapter 14 (weighting/estimation) of the National Survey Design and Planning Report (NSDPR). Countries were required to complete and return the forms at least six months prior to the start of the Field Test data collection. This deadline was set to ensure Field Test sample design and selection steps provided all the necessary opportunities to test various aspects of the Main Study sample design and selection activities. Countries then had the opportunity to update their Main Study plans after the Field Test.

Sampling Plan Form Part 1 addressed the standards and guidelines related to sample design and selection. It was to be completed separately for the Field Test and Main Study. The form included questions on country plans for each of the following:

- Country-Specific Supplemental Samples
- Target Population Definition
- Background Design Information
- Sample Design and Sampling Units
- Within-Household Selection Rule (for countries with DU sampling)
- Sampling Frame Description
- Coverage Rate of Target Population
- Sample Selection Methods for Area Units (if applicable)
- Sample Selection Methods for DU and Within-Household Sampling (if applicable)
- Sample Selection Methods for Persons from Registries (if applicable)
- Sample Selection Checks
- Pre-Assignment of Assessment Instruments
- File Delivery
- Initial Sample Size Worksheet
- Reserve Sample
- Data Consistency Checks
- Sample Monitoring Plans
- Incentives

Sampling Plan Forms Part 2 and Part 3 pertained to the Main Study only. Part 2 checked countries' ability to comply with the weighting chapter (Chapter 14) of the PIAAC Technical Standards and Guidelines. It included questions on potential variables for weighting adjustments, planned weighting procedures, and the intended variance estimation method. Part 3 addressed expected response rates and NRBA plans.

Sample selection quality control forms

The QC sample selection (SS) forms collected detailed information about the country sample selection process and the results. Countries were to submit forms after each sample selection stage, allowing adequate time for countries to respond to the Consortium comments and questions and to revise procedures if necessary. The forms were important to verify that the selection of a probability sample adhered to the PIAAC Technical Standards and Guidelines.

The forms covered the following:

- Definition of the sampling unit
- Variables used for stratification, sorting and measure-of-size calculations
- List of certainty units, such as large primary sampling units
- Average, minimum and maximum cluster size
- Number of units on the frame, number of units sampled, weighted totals and target population totals, by stratum
- Weighted population totals by characteristics of interest (such as region or age)

- Weight distributions, where the weight is the inverse of the selection probability
- Description of any oversampling

Sample monitoring quality control forms

The sample monitoring process was intended to help countries identify potential shortfalls in the sample, problems in achieving the desired response rate, and the potential for nonresponse bias in the collected sample. Continuous monitoring was used to allow countries to employ procedures to address these problems during data collection while it was still possible to meet goals associated with sampling and data quality. Countries were required to complete QC sample monitoring (SM) forms every one to two months during data collection. The Consortium reviewed the forms and provided feedback to countries. The SM-1 forms collected information by key subgroups on the number of cases completed, response rates and expected yield. Countries were asked to monitor these figures by gender, age groups, geography and other characteristics of interest in order to help identify any shortfalls in yield or unusually low response rates. Starting mid-data collection, countries were also asked to provide a more extensive NRBA (SM-2) to identify subgroups with low response rates. The subgroups could be formed according to demographic or area-level characteristics believed to be related to proficiency. Multivariate techniques, such as a classification tree algorithm, were recommended for this evaluation to identify subgroups created from combinations of key variables.

Sampling-related quality control data checks

The Consortium provided countries with suggested sampling-related QC checks that the countries could run during data collection. These checks were intended to supplement the record consistency checks in the Data Management Expert (DME) software and emphasized variables relating to the Sample Design International File. Instructions were provided for checking consistency among disposition codes at the screener level (if applicable) and background questionnaire (BQ) level, checking the sampling of persons, and reviewing the conditions for a completed case as defined in standard 4.3.3.

Sample Design International File (SDIF) and Weighting International Files (WIFs)

At the end of data collection, countries provided the Consortium with an SDIF that contained sample selection data for each sampled unit, including sampling strata, probabilities of selection, ID variables, disposition codes, and auxiliary variables for weighting adjustments. The SDIF was the input file to the weighting process. The Consortium performed QC checks on the file to verify that variable definitions and formats were consistent with the specifications in Annex 4-3 of the PIAAC Technical Standards and Guidelines and that those fields reflected the information provided by the countries in their sample selection forms and weighting plans.

Countries also provided WIFs for Benchmark Control Totals to the Consortium. The files contained the external control totals to be used in the benchmarking adjustments. The benchmark WIFs were reviewed to check that the overall target population total was the same for each variable used in the benchmarking adjustment and that there was a set of control totals for each benchmarking variable included on the SDIF. Countries performing their own weighting adjustments also supplied a WIF for Quality Control Checks that was used to supplement the checks performed through the weighting QC forms (described below).

So as to not jeopardize the weighting schedule due to data reconciliation issues, countries were asked to provide a preliminary version of the SDIF and benchmark WIF before the end of data collection.

Weighting quality control forms

The Consortium developed a set of QC checks to review the weighting process for adherence to the weighting standards and guidelines and to check weight calculations for reasonableness and accuracy. Prior to the weighting period, each country needed to complete and return a checklist on the PIAAC Technical Standards and Guidelines related to weighting (Weighting QC Form W-0). They indicated whether the standards and guidelines were consistent with their implementation and understanding and indicated any deviations. They also needed to complete a W-1 form that contained checks on the base weights, variance strata and variance unit assignments, and any imputation performed for weighting variables.

Countries could opt to have the Consortium perform the weighting adjustments, or they could choose to create the final sampling weights themselves. During weighting, countries that formed their own weights were required to report on details of their weighting adjustments and weight distributions through a series of QC forms. If the Consortium conducted the weighting steps, the Consortium provided the forms to the countries for their review.

Form W-2 covered the household weights for countries with a household stage of sampling. Form W-3 was on the person-level weighting adjustments, and Form W-4 dealt with the final weights. The forms included the following checks:

- Descriptive statistics (including the counts of cases with missing and nonmissing weights, and sum, mean, minimum, maximum, and CV¹ of weights) on the full sample weights across weighting stages for all the sample, and by region, age group, and gender respectively
- Sum of replicate weights across weighting stages
- Descriptive statistics on selected replicate weights across weighting stages
- Unweighted and weighted counts by response status and weighting adjustment cells across weighting stages
- Description of trimming procedures
- Listing of the largest weights
- Comparison of control totals to external totals and weighted PIAAC totals
- Design effect calculations

Performing the weighting QC checks was essential for verifying that the final weights produced for estimation were appropriate. If any issues with the weighting adjustments were identified by the weighting QC forms, countries were required to rectify the problems and resubmit the QC forms until no more issues were found.

¹ Refer to section 16.4.2 for the definition of CV.

Weighted response rates and NRBA

Regardless of response rate, all countries were required to conduct a basic NRBA. The basic analysis evaluated the relationship of response status to available auxiliary variables and provides an indication of nonresponse bias prior to weighting adjustments. It could be used to inform the choice of weighting variables.

As described in section 16.2, the Consortium computed weighted response rates for each country using the official response rate formulae in Annex 4-3 of the PIAAC Technical Standards and Guidelines and the data provided on the countries' SDIF. If a country's overall response rate fell below 70%, or if it had a stage of data collection with a response rate of less than 80%, the country was then asked to conduct an extended NRBA. This analysis included the evaluation of the potential for remaining bias after weighting adjustments were completed. It also attempted to evaluate bias directly in the proficiency estimates rather than solely relying on auxiliary variables.

Finally, countries were required to compute item response rates and conduct an item nonresponse bias analysis for any BQ items with response rates below 85%. The analyses were similar to those for the basic unit NRBA and involved comparing characteristics of item respondents and nonrespondents.

16.2 Sampling coverage and response rates

Coverage rates and response rates are important measures of the quality of the survey because they reflect the representation of the target population. Countries focused on reducing noncoverage and nonresponse bias given that the main goal of PIAAC is to produce high-quality unbiased estimates of the target population that are comparable across countries. First, section 16.2.1 contains an introduction to the implications of noncoverage and nonresponse on the potential for bias in the survey results. This will be discussed further in section 16.3. Then we turn to the computation of the coverage rates and the response rates.

16.2.1 Potential for bias

Under ideal situations, every eligible adult in the target population would have a nonzero chance of selection in a national sample, would be located and would agree to participate in the study. In practice, these circumstances are not realized in any survey population. There is a potential for bias whenever part of the target population is excluded from the frame or sampled persons who did not participate in the survey have different characteristics than those who did. For some important characteristics, the respondents may be substantially different from the rest of the target population, resulting in biased outcome estimates.

When response rates are low, there is a greater chance for nonresponse bias. The extent of nonresponse bias depends on how correlated the response propensity is with the survey outcomes. It is, therefore, critical to evaluate the potential for nonresponse bias, as a quality check on the estimates, at the conclusion of the data collection. Similarly, noncoverage bias (due to exclusions) can be substantial if the noncoverage rate is high and the difference in proficiency levels between adults included in the sample and those excluded from the frame is relatively large. Given the relationships between bias and coverage and response rates, countries had to

keep the exclusion rates low and implement procedures to reduce the potential for nonresponse bias and attain high response rates.

The maximum allowable exclusion rate was set at 5% to guard against high noncoverage bias in PIAAC estimates. Any exclusions to the core PIAAC target population, whether or not they exceeded the threshold, were reviewed and approved by the Consortium. Even though up to 5% exclusions were tolerated, exclusions had to be kept to a minimum. If the quality of the sampling frame was such that it could result in a noncoverage rate of more than 5%, participating countries had to look into ways to improve coverage.

To reduce the potential for nonresponse bias, countries had to plan and implement field procedures that obtain a high level of cooperation. It was critical to monitor the distribution of the sample during data collection to ensure steps were taken to reduce the potential for bias as much as possible. As nonresponse rates increased, countries actively had to seek auxiliary data to reduce the impact of response propensities on the survey estimates. These auxiliary variables were used in weighting adjustments for the purpose of reducing nonresponse bias. Although sample weight adjustments based on auxiliary data are effective in reducing nonresponse bias, they are not considered as replacements for a vigorous effort to achieve the highest response rate possible.

16.2.2 Coverage rates

The PIAAC target population is defined as all noninstitutionalized adults between the ages of 16 and 65 (inclusive) who reside in the country at the time of data collection. The PIAAC Technical Standards and Guidelines require that the sampling frame covers at least 95% of the PIAAC target population. Exclusions (that is, persons who had no chance of being selected into the sample) may represent no more than 5% of the target population. There are, in effect, two categories of exclusions in PIAAC – *ex ante* exclusions by design (frame exclusions) and *ex post* exclusions following data collection (inaccessible persons). Both contribute to the overall noncoverage rate.

Exclusions by design

Exclusions by design or frame exclusions are of two types. They include, first, exclusions resulting from a decision not to include certain population groups in the sampling frame (e.g., the populations of remote and isolated regions) for reasons such as difficulty of access and the resulting high cost of data collection. Second, the use of a particular sampling frame may lead to the exclusion of certain groups in the population by virtue of the rules that determine which individuals are included in the list constituting the frame. For example, many population registers include only those members of the population with valid residence permits and, therefore, exclude illegal immigrants.

The frame noncoverage rate is computed as the estimated population in the excluded groups divided by the estimated core PIAAC target population. The rates by country are provided in Table 16-2. More information on sampling frame noncoverage, including the specific groups excluded by each country, is provided in Chapter 14.

Exclusions related to data collection

In addition to persons who are eligible under the international definition of PIAAC target population but were not included in the frame, persons that were included in the frame but in practice were impossible to be interviewed could be treated as exclusions. Some registry-based countries experienced difficulties locating and interviewing some or all sampled persons not residing at the address listed in the registry. Such cases were classified into a number of categories, as shown in Table 16-1. To arrive at an optimum and consistent approach across all registry-based countries, the Consortium assumed that all countries tried to find the location of the sampled persons and tried to interview them if they moved into one of the PSUs in the sample or were in a location where it was possible for PIAAC interviewers to visit and conduct the interview and assessment. Some individuals are found to be out of scope when the contact is attempted (e.g., information is provided that indicates that they have died, moved to an institutional setting, or emigrated). Others are “inaccessible” in that they cannot be interviewed because the information about their residential address was incorrect or because they have moved to another location in the country, which means they cannot be interviewed. Finally some members of the sample are untraceable in that no information about their whereabouts is available. The main advantage of classifying such cases in this manner was that the information about the inaccessible cases could be used to reduce the bias associated with noncoverage and, thus, reduce inconsistencies between country data.

The inaccessible noncoverage rate was calculated as the inaccessible population divided by the eligible population. The observed noncoverage rate had to incorporate sampling weights to account for selection probabilities and to ensure that the observed rate was representative of inaccessibles in the frame. If countries had an overall noncoverage rate (including frame and inaccessibles) of greater than 5%, up to 5% were reported in the noncoverage rate and the portion greater than 5% contributed as nonresponse in the response rate calculations.²

Table 16-2 shows the noncoverage rates for each country.

Table 16-1: Registry-based samples: Categories of ‘non-contacts’ and their status

Description	Status
Deceased	Out of scope
Moved outside country	Out of scope
Moved inside country	
Moved into institution	Out of scope
To PIAAC PSU	Inaccessible (unknown or invalid address)
To non-PIAAC PSU	Inaccessible (inability to interview outside PIAAC PSUs)
To unknown PSU	Inaccessible
Unknown whereabouts	Distributed between “out of scope” and “inaccessible” categories
Invalid address	Inaccessible

² This differs from the treatment of inaccessibles in weighting. For weighting purposes, such cases were treated as nonrespondents (see Chapter 15).

Table 16-2: Noncoverage rates: Sampling frame and inaccessible within sample

Country	Noncoverage Rate		
	Sampling Frame	Inaccessible	Overall
Australia	3.3%	0.0%	3.3%
Austria	0.6%	0.8%	1.4%
Canada	1.8%	0.0%	1.8%
Cyprus ³	<2.0%	0.0%	<2.0%
Czech Republic	1.8%	0.0%	1.8%
Denmark	<0.1%	5.0%	5.0%
England (UK)	2.0%	0.0%	2.0%
Estonia	2.8%	0.6%	3.4%
Finland	0.2%	0.5%	0.7%
Flanders (Belgium)	1.0%	4.0%	5.0%
France	<2.6%	1.4%	<4.0%
Germany	0.5%	2.0%	2.5%
Ireland	0.4%	0.0%	0.4%
Italy	0.8%	1.9%	2.7%
Japan	2.2%	2.8%	5.0%
Korea	2.4%	0.0%	2.4%
Netherlands	0.9%	1.8%	2.7%
Northern Ireland (UK)	2.0%	0.0%	2.0%
Norway	0.4%	0.4%	0.8%
Poland	1.0%	4.0%	5.0%
Russian Federation ⁴	1.5%	0.0%	1.5%
Slovak Republic	0.1%	4.9%	5.0%
Spain	0.0%	5.0%	5.0%
Sweden	<1.0%	0.0%	<1.0%
United States	0.1%	0.0%	0.1%

16.2.3 Response rates

Response rate is a valuable data quality measure and the most widely used indicator of survey quality. A high response rate increases the likelihood that the survey accurately represents the target population, and a low response rate reflects the possibility of bias in the outcome statistics.

A minimum overall response rate of 70% was set as the goal for PIAAC countries to be included in international indicators and reports, unless sample monitoring activities and/or nonresponse bias analyses indicate serious levels of bias in the country data. Countries with response rates of between 50% and 70% were included in international indicators and reports, unless other factors like noncoverage bias were detected. Deviations from the international standards on response rates were, however, documented in the international reports and publications. Results from countries with response rates below 50% were not published unless the country provided the

³ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

⁴ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

OECD Secretariat with evidence that the potential bias introduced by the low response rates was unlikely to be greater than the bias associated with response rates of between 50% and 70%.

Using the standard formulae shown in Table 16-3, weighted response rates were computed hierarchically for the following stages of data collection:

- Screener (if the sample design included a screener stage)
- Background questionnaire
- Assessment (without and without reading components)
- Overall

Table 16-3: Response rate

Stage	Response Rate Calculation	Description
Screener	COMPLETE / ELIGIBLE COMPLETE = C^s ELIGIBLE = $HH^s - I^s - U^s * (I^s / K^s)$	C^s = Completed screeners, HH^s = All sampled households, I^s = HHs known to be ineligible, U^s = HHs with unknown eligibility status, K^s = HHs with known eligibility status.
Background Questionnaire (For countries with screeners)	COMPLETE / ELIGIBLE COMPLETE = $C^b + LR^b$ ELIGIBLE = $SP^b - D^b - I^b$	C^b = Completed BQ cases, LR^b = Literacy-related nonrespondents, SP^b = All sampled persons, D^b = SPs with a disability, I^b = SPs known to be ineligible.
Background Questionnaire (For countries with registries)	COMPLETE / (ELIGIBLE – EXCLUDE) COMPLETE = $C^b + LR^b$ ELIGIBLE = $SP^b - D^b - I^b - U^b * ((D^b + I^b) / K^b)$ EXCLUDE = ELIGIBLE * EXC_PROP	C^b = Completed BQ cases, LR^b = Literacy-related nonrespondents, SP^b = All sampled persons, D^b = SPs with a disability, I^b = SPs known to be ineligible, U^b = SPs with unknown eligibility status, K^b = SPs with known eligibility status. EXC_PROP = Inaccessible rate from Table 16-2
Assessment ¹	COMPLETE / ELIGIBLE COMPLETE = $C^a + LR^a$ ELIGIBLE = $C^b - D^a - I^a$	C^a = Completed assessments, LR^a = Literacy-related nonrespondents, C^b = Completed BQ cases, D^a = SPs with a disability, I^a = SPs known to be ineligible.

¹ The assessment response rates with and without reading components were computed using the same formula, the difference being reflected in how each SP was classified, whether completing the reading components or not.

The literacy-related cases were included in the numerator of the response rates because their reason for nonresponse provides an indication of their proficiency level. The disabilities, while considered in scope, were subtracted from the denominator because the assessment did not accommodate such situations.

Table 16-4 shows a summary of the response rates for the participating countries.

Table 16-4: PIAAC response rates for participating countries

Country	Reading component	Response Rates					
		Without Reading Component				With reading component	
		Screeners	BQ	Assessment	Overall	Assessment	Overall
Australia	Yes	85%	88%	96%	71%	96%	71%
Austria	Yes	.-	53%	99%	53%	99%	53%
Canada ¹	Yes				59%		58%
Cyprus ⁵	Yes	74%	99%	100%	73%	100%	73%
Czech Republic	Yes	74%	90%	100%	66%	100%	66%
Denmark	Yes	.-	51%	97%	50%	97%	50%
England (UK)	Yes	89%	68%	97%	59%	97%	59%
Estonia	Yes	.-	64%	99%	63%	99%	63%
Finland	No	.-	69%	95%	66%	.-	.-
Flanders (Belgium)	Yes	.-	62%	99%	62%	99%	62%
France	No	.-	71%	94%	67%	.-	.-
Germany	Yes	.-	55%	99%	55%	100%	55%
Ireland	Yes	79%	92%	99%	72%	99%	72%
Italy	Yes	88%	66%	97%	56%	97%	56%
Japan	No	.-	50%	100%	50%	.-	.-
Korea	Yes	86%	91%	96%	75%	96%	75%
N. Ireland (UK)	Yes	83%	80%	98%	65%	98%	65%
Netherlands	Yes	.-	53%	97%	51%	98%	51%
Norway	Yes	.-	63%	98%	62%	98%	62%
Poland	Yes	.-	56%	99%	56%	95%	54%
Russian Federation ⁶	No	53%	99%	97%	52%	.-	.-
Slovak Republic	Yes	.-	66%	99%	66%	99%	66%
Spain	Yes	.-	48%	100%	48%	100%	48%
Sweden	Yes	.-	46%	97%	45%	97%	45%
United States	Yes	86%	83%	99%	70%	99%	70%

¹ To account for multiple sampling frames and to provide an indication of nonresponse bias, nonresponse to the parent samples were reflected in Canada's PIAAC overall response rate computation. (See Chapter 14 for information on Canada's sample design.) It was decided that individual response rates at the screener, BQ and assessment stages are not to be reported.

⁵ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

⁶ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Some countries chose to conduct the optional reading components as part of the psychometric assessment, and response rates with reading components were calculated for them. For all countries response rates were calculated without reading components, which provide a comparable measure across the countries. For countries with a screener, the overall response rate was calculated as the product of the response rates for the screener, BQ and assessment. For countries without a screener, the overall response rate was calculated as the product of the response rates for the BQ and the assessment. The screener response rate was weighted by the inverse of the household selection probability, and the BQ and assessment response rate by the inverse of the person selection probability. If countries had oversampling, it is reflected in the weights, and therefore weighted response rates are a comparable measure across countries.

16.3 Nonresponse bias analysis

Missing data can occur when some of the adults selected in the sample are not contacted or refuse to participate (referred to as unit nonresponse), they fail to respond to a particular survey item (referred to as item nonresponse), or because data collected from the sampled adults is contaminated (and thus not useful) or lost during or after the data collection phase. Nonresponse bias can be substantial when two conditions hold: 1) the response rate is relatively low and 2) the difference between the characteristics of respondents and those of nonrespondents is relatively large. This is reflected in the following deterministic nonresponse bias formula:

$$Bias(\bar{y}_R) = (1 - W_R)(\bar{Y}_R - \bar{Y}_{NR}),$$

where W_R is the proportion of respondents, \bar{Y}_R is the mean outcome for respondents, and \bar{Y}_{NR} is the mean outcome for nonrespondents. An alternative model of nonresponse assumes each sampled person has a certain propensity to respond, and nonresponse bias in a characteristic is a function of the covariance between the response propensity and the characteristic:

$$Bias(\bar{y}_R) = \frac{\sigma_{yp}}{\bar{p}},$$

where σ_{yp} is the covariance between the outcome variable and response propensity, and \bar{p} is the mean response propensity. Based on this model, NRB is present if missingness is related to proficiency, as measured by PIAAC.

Countries worked to reduce nonresponse bias to the extent possible before, during, and after data collection. Before data collection, countries implemented field procedures with the goal of obtaining a high level of cooperation. Most countries followed the PIAAC required sample monitoring activities to reduce bias to the lowest level possible during data collection. Finally countries gathered and used auxiliary data to reduce bias in the outcome statistics through nonresponse adjustment weighting.

All countries were required to conduct a basic nonresponse bias analysis (NRBA) and report the results. The basic analysis was used to evaluate the potential for bias and to select variables for nonresponse adjustment weighting. In addition, countries were required to conduct and report the

results of a more extensive NRBA if the overall response rate was below 70%, or if any stage of data collection (screener, background questionnaire, or the assessment) response rate was below 80%. An item NRBA was required for any BQ item with response rate below 85%.

A summary of the results of the basic NRBA is provided in Section 16.3.1. Section 16.3.2 contains the results of the extended NRBA, and Section 16.3.3 provides a summary of the item nonresponse analysis. A brief summary and conclusions of the NRBA is given in Section 16.3.4.

16.3.1 Basic NRBA

The basic NRBA involved comparing survey respondents and nonrespondents using auxiliary variables available on the sampling frame, available from a previous data collection stage (e.g. screener data for the BQ analysis), or coming from an external source that could be matched to each sampled unit. Also, observational data on respondents and nonrespondents collected during data collection could have been used to evaluate bias, assuming the data was of sufficient quality. The auxiliary variables must have been available for all eligible units and, as noted above, had to be related to proficiency. All countries were required to include the following variables in their analysis: age, gender, education, employment, and region. If any of these variables was not available for all eligible units, then a corresponding area-level variable could have been used instead (e.g. the employment rate within small geographic areas).

The basic analysis included results from the following:

- Comparison of response rates for different subgroups
- Use of a chi-square test or estimates of relative bias to compare the distribution of auxiliary variables (correlated with proficiency) for respondents and nonrespondents
- Use of a classification tree algorithm to identify subgroups with low response rates or use of logistic regression to model the relationship between response status and the auxiliary variables

The response rate and chi-square analyses were useful in explaining the relationship of response status to each auxiliary variable individually. A classification tree algorithm and/or a logistic regression model was used to evaluate the relationship between response status and multiple auxiliary variables.

All countries completed all the required analyses and included all the required variables, age, gender, education, employment, and region, in their analysis, with the exception of Austria, Finland, Flanders (Belgium) and Italy. In most cases, the failure to include the required variables in the analyses was due to the lack of access to sources with reliable data for such variables.

An initial basic NRBA was conducted prior to the weighting process. The analysis was conducted in two stages. The first stage helped to create a pool of predictor variables related to proficiency, using the field test data. The second stage helped to reduce the pool of predictor variables to those related to response propensity (this was repeated after the weighting process to finalize the basic NRBA). Most countries used all auxiliary variables that showed potential for bias in deriving nonresponse adjustments to the sampling weights. The remaining countries used most of the variables identified in the initial basic NRBA, mainly because reliable data was not available for the remaining variables.

Nonresponse weighting adjustments reduce bias in the outcome statistics to the extent that auxiliary variables are correlated with proficiency. Mainly, weighting adjustments are carried out by assuming nonrespondents' proficiency levels are the same as the respondents in the subgroups created for weighting adjustments using the auxiliary variables. This assumption is, of course, not true and the level of bias reduction depends on the number of auxiliary variables used during weighting and the correlation between these variables and proficiency.

The basic NRBA is a good initial assessment of nonresponse bias and is essential in identifying effective weighting variables. However, it has its limitations. The analysis does not reflect the effect of weighting adjustments on NRBA, and the extent of bias remaining after nonresponse adjustments are conducted. Therefore, countries with lower response rates were required to conduct a more extensive analysis to assess the potential for bias remaining after nonresponse adjustment weighting. Section 16.3.2 includes a brief description of the results of the extended NRBA.

16.3.2 Extended NRBA

A more extensive NRBA was required if the overall response rate was below 70%, or if any stage of data collection (screener, background questionnaire, or the assessment) response rate was below 80%.

Australia, Korea and the United States achieved an overall response rate of 70% or greater, with response rates for each stage being greater than 80%, and thus did not require the extended NRBA. Cyprus⁷ and Ireland also achieved overall response rates of 70% or greater, but they achieved a lower than 80% response rate for one stage of their sample. The remaining countries achieved response rates lower than 70%.

The main purpose of the extended analysis was to assess potential for remaining bias in the final weighted proficiency estimates after adjusting for nonresponse. Because the proficiency levels of nonrespondents are unknown, the NRBA is carried out by making assumptions about nonrespondents. Therefore, it is necessary to conduct multiple analyses to assess the potential for bias since each analysis has its own limitations resulting from the specific assumptions made about nonrespondents. The extended NRBA included seven analyses (as listed below). Together, they were used to assess the patterns and potential for bias in each country data.

The extended NRBA included the following analysis:

1. Comparison of estimates before and after weighting adjustments;
2. Comparison of weighted estimates to external totals;
3. Correlations of auxiliary variables and proficiency estimates;
4. Comparison of estimates from alternative weighting adjustments;
5. Analysis of variables collected during data collection;

⁷ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

6. Level-of-effort analysis; and
7. Calculation of the range of potential bias.

These analyses are described further below.

Cyprus⁸ and Ireland were required to do only a subset of the analysis since their overall response rate was higher than 70%.

Comparison of estimates before and after weighting adjustments

To better capture the effects of the weighting adjustments on unit nonresponse bias, estimates from the full sample were compared to estimates from the respondents before and after weighting adjustments. To compare estimates before and after each step of weighting adjustments, the following comparisons were made:

- Comparison of percentage distributions from BQ base weights for the total eligible sample of persons with the BQ base weights for the BQ respondents to check for differences due to nonresponse to the BQ
- Comparison of percentage distributions from BQ base weights for the total eligible sample of persons with that from the BQ nonresponse adjusted weights for respondents to check for differences after the nonresponse adjustment process to the BQ
- Comparison of percentage distributions from BQ nonresponse adjusted weights for respondents with that from the BQ raked weights (weights adjusted to two or more marginal population totals) for respondents to check for differences that may have been introduced through the initial raking procedure

For countries that had screeners, analogous comparisons to the BQ level, as mentioned above, were completed. All the countries required to do the analysis completed it. The goal was to include at least one auxiliary variable not present in weighting adjustments in addition to those used during nonresponse adjustment weighting. Inclusion of the non-weighting variables shows whether the weighting adjustment was effective in reducing bias in other known auxiliary variables, not just the weighting variables. The following 11 countries; Denmark, England (UK), Finland, Germany, Japan, Netherlands, Northern Ireland (UK), Norway, Poland, Slovak Republic and Sweden and included nonweighting variables in this analysis as well as weighting variables. The remaining countries only included the weighting variables. Canada included a substantial number of weighting variables in their analysis. In general, all countries except for Russian Federation (partial compliance) observed that bias was reduced in auxiliary variables through weighting adjustments.

Comparison of weighted estimates to external totals

The second analysis compared estimates from PIAAC to external source estimates to assess potential for bias in PIAAC outcome statistics.

To the extent possible, countries used estimates from external sources that measured the same characteristic for a similar time period. Some external source estimates were subject to sampling

⁸ See above footnote.

error also, and thus the variance of these estimates were taken into account when making comparisons.

Many countries found significant differences between the PIAAC estimates and the external source estimates but were mostly able to explain the sources for discrepancies. The sources mainly included, different data collection time periods or different definitions (e.g., definition of employment). All countries except France completed this analysis.

Correlations of auxiliary variables and proficiency estimates

The analyses described thus far relied on auxiliary variables and did not directly measure bias in the proficiency estimates. Bias in the auxiliary variables is indicative of bias in the proficiency estimates to the extent that the auxiliary variables and proficiency estimates are correlated. Thus, correlations between the auxiliary variables and proficiency data are good indicators of potential for bias reduction through weighting adjustments. For variables used in the weighting adjustments, a low correlation with proficiency implies that using the variable in the weighting adjustments did little to reduce nonresponse bias. On the other hand, a high correlation with proficiency implies a potentially high reduction in nonresponse bias. However, it should be noted that the correlations are based on respondents' data, and the relationship between proficiency and the auxiliary variables might be different for nonrespondents. Therefore, the correlations could be different if a country's response rate is very low, and if nonrespondents are different from respondents in terms of the relationship between their scores and the auxiliary variables.

Correlations were calculated as the square root of R-square of a weighted analysis of variance, whose dependent variable was the literacy or numeracy score while the explanatory variables were the weighting variables (BQ nonresponse adjustment cells and raking dimensions).

Table 16-5 presents the correlation between the proficiency and the weighting variables for each country.

Table 16-5: Correlations of auxiliary variables and proficiency estimates

Country	Literacy	Numeracy
Austria	0.56	0.57
Canada	0.54	0.53
Cyprus ⁹	0.39	0.47
Czech Republic	0.56	0.60
Denmark	0.50	0.46
England (UK)*	0.52	0.56
Estonia	0.37	0.35
Finland	0.60	0.58
Flanders (Belgium)	0.36	0.36
France	0.60	0.64
Germany	0.61	0.62
Ireland	0.52	0.53
Italy	0.49	0.53

⁹ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 16-5 (cont.): Correlations of auxiliary variables and proficiency estimates

Country	Literacy	Numeracy
Japan	0.53	0.52
Korea	0.55	0.55
Netherlands	0.57	0.55
Northern Ireland (UK)*	0.57	0.60
Norway**	0.48	0.48
Poland	0.40	0.37
Russian Federation ¹⁰	0.35	0.34
Slovak Republic	0.38	0.38
Spain	0.62	0.62
Sweden	0.70	0.70
United States	0.63	0.66

*England (UK) and Northern Ireland (UK) were weighted separately to allow efficient estimates for each population.

** Norway was not able to provide nonresponse adjustment cells due to confidentiality concerns. Therefore, Norway self-reported the correlation between literacy scores and BQ nonresponse adjustment variables and raking variables as 0.48 for literacy. Norway did not report the correlation for numeracy. Therefore, 0.48 was assumed for numeracy.

There are a few countries with low correlation between the BQ nonresponse cells and the proficiency scores. However, all of the correlations between proficiency scores and the BQ nonresponse cells and the raking dimensions combined are higher than 0.30 and the average is 0.51 for literacy scores and 0.52 for numeracy scores. Although it was not required, the correlations for Korea and the US were also provided. Based on the moderate-to-high correlations between the weighting variables and the proficiency scores, we can expect the weighting adjustment to have reduced bias in the proficiency scores.

Figure 16-1 displays each country's correlation between weighting variables and the literacy score and correlation between weighting variables and the numeracy score. The two correlations are very close to each other, implying the same level of effectiveness in reducing bias for the two proficiency estimates.

¹⁰ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Figure 16-1. Correlation of weighting variables and the proficiency scores

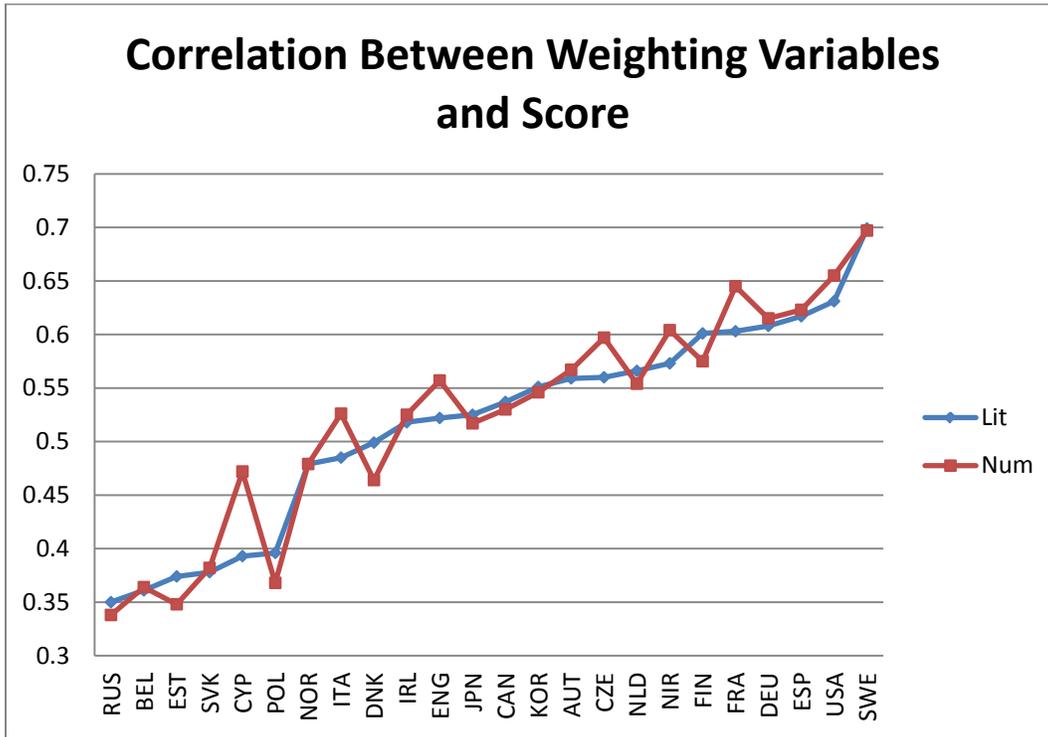


Figure 16-2 shows the plot of response rate versus correlation between the weighting variables and the literacy score reflecting the effectiveness of nonresponse adjustments in reducing bias.

Figure 16-2. Scatterplot of response rate versus correlation

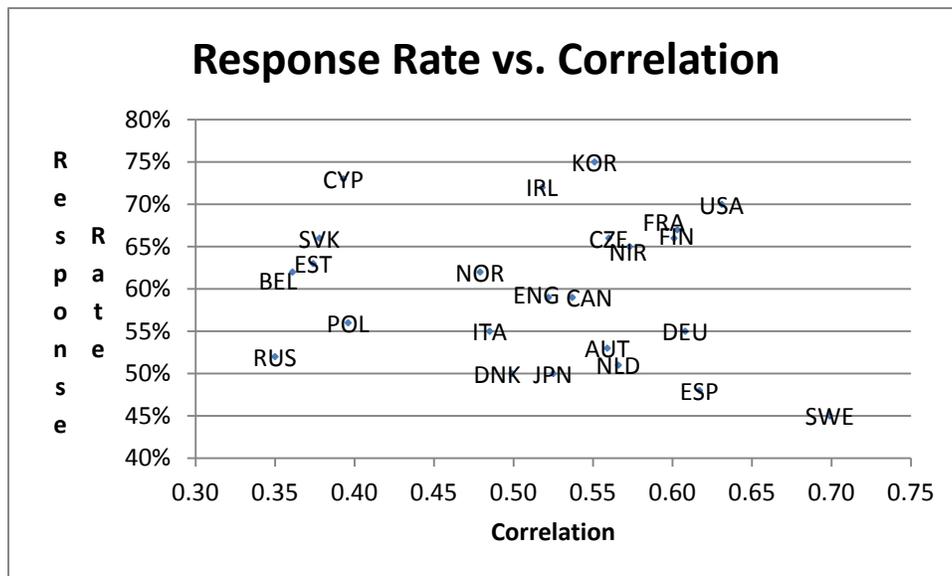


Figure 16-2 shows that:

- Countries in the lower right corner, such as Sweden, Spain and Germany, have low response rates, but are expected to have accomplished a considerable bias reduction through weighting, since their weighting variables are highly correlated with the proficiency.
- Austria, Canada, Denmark, England (UK), Italy, Japan and Netherlands have about average correlations, so bias reduction is expected at an average level as compared to other countries.
- Finland, France and the US have a higher than average correlation and high response rates.
- Cyprus,¹¹ Estonia, Flanders (Belgium) and Slovak Republic have low correlations, but relatively high response rates, which helped reduce potential for bias. Poland and Russian Federation, which also have low correlations, have somewhat lower response rates, which indicates relatively less potential for bias reduction.

Comparison of estimates from alternative weighting adjustments

For this evaluation, an auxiliary variable was re-calibrated to known totals, and estimates of the key statistics were compared before and after the re-weighting. Re-weighting was useful as an evaluation tool when:

- The variable was not used in weighting (because it was not available) or was used but with different categories
- The variable is correlated with the outcome measure
- The variable is correlated with response propensity

Any differences between estimates using the official survey weights and the re-weighted weights reflected noncoverage as well as nonresponse bias, but if there was not a large change in the estimates, this was further confirmation that nonresponse bias may not be a concern.

Thirteen of the countries fully complied with the analysis and results confirmed that nonresponse bias may not be a concern. These countries were: Austria, Canada, Denmark, Estonia, Finland, Flanders (Belgium), Germany, Japan, Netherlands, Norway, Poland, Spain and Sweden. Italy found a significant difference between the average literacy score using final weights and when using the alternative weights, where the alternative weights were created using a more detailed weighting variable. Some caution should be used in conclusions from this analysis for Czech Republic (quality unknown due to unavailability of data), France (did not comply), Russian Federation (did not comply), Slovak Republic (partial compliance) and UK (did not comply).

Japan and Sweden used the results of this analysis to improve their final survey weights.

Analysis of variables collected during data collection

Disposition codes contain information on reasons for nonresponse. For this analysis, distributions of sampled persons with known characteristics related to outcome (i.e. the literacy-related nonrespondent (LRNR) cases, which are language problems, reading and writing difficulty, and mental disability) were examined. For example, the demographic distribution of

¹¹ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

literacy-related cases was compared to other eligible persons using auxiliary data, and interview data. Statistical tests such as Chi-square tests were processed to determine if there is a relationship between select demographic variables and the disposition codes for nonrespondents. A special weighting adjustment for literacy-related cases was conducted for all countries, with the exception of Poland, where the BQ LRNRs together with the other BQ NRs were represented by BQ respondents. Therefore, in almost all countries, the existence of LRNR cases was dealt with appropriately in order to reduce potential for bias.

All countries, except for France, conducted an analysis of disposition codes with some observing differences that were expected, given the conditions in their countries. However, Sweden and the UK each conducted only a partially completed analysis (i.e., the quality level is unknown) due to unavailability of data.

In addition, Non-Interview Report (NIR) forms identify observable demographic information and reasons for nonresponse that are not captured in the disposition codes. The NIR forms can potentially indicate whether the reasons for nonresponse are related to proficiency estimates and suggest ways to improve response rates for future surveys.

The following countries put extra effort in conducting the analysis using the information from NIR forms: Cyprus¹², Germany, Italy, Japan, and Slovak Republic. The observed information from NIR forms may be useful for data collection in the next cycle.

Level-of-effort analysis

Another way to evaluate bias in the proficiency estimates is to compare proficiency estimates by level of effort. To the extent that the late or hard-to-reach respondents are similar to the nonrespondents, differences in proficiency estimates between the late and early (or hard-to-reach and easy-to-reach) respondents could indicate nonresponse bias. This analysis can be useful in detecting potential for bias given the assumption that nonrespondents are similar to respondents at the end of the data collection period.

If the literacy estimates differed between easy and hard respondents within a category of a weighting variable (used in the level-of-effort analysis), that may indicate that there are differences even within the weighting cells, and the nonresponse adjustment might not have helped. However, it may be that the data collection procedures were effective in obtaining a different type of respondent, potentially reducing the bias.

Thirteen countries revealed some significant differences in characteristics between early and late respondents, including Austria, Cyprus¹³, Czech Republic, Denmark, Estonia, Flanders (Belgium), Italy, Japan, Netherlands, Norway, Poland, Spain and Sweden. Two countries, Finland and Germany, conducted the analysis but did not find significant differences. France, Russian Federation (due to the inability to classify respondents as difficult-to-contact) and Slovak Republic did not comply with the analysis, and some caution should be used in drawing conclusions from UK's analysis due to unavailability of data.

¹² Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

¹³ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Calculation of the range of potential bias

The final component of the bias analysis is to evaluate the potential for bias remaining after weighting under the scenario that nonrespondents' proficiency scores are vastly different from the assumptions made during weighting.

It is well known that NRB can be reduced to some unknown extent through sample weighting when proficiency is correlated with auxiliary variables, and auxiliary variables are correlated with response propensity. Weighting assumes response probabilities are constant within every group created for weight adjustment, the proficiency score has zero variance within each group, and response propensity is uncorrelated with proficiency. It is known that these assumptions are not correct, and the impact of weight adjustments is limited to the number of variables available for nonresponse adjustment, and correlation levels with proficiency. Also, it is not possible to measure the exact departure from these assumptions since proficiency levels of nonrespondents are not known. This analysis attempts to evaluate the potential for bias by computing a range based on an extreme assumption that nonrespondents would all score at the 10th percentile within each weighting cell, and at the other extreme they would all score at the 90th percentile within each weighting cell. The range of bias was computed as the difference between the two extreme estimates, while taking into account the response rate and population size in the weighting cell.

The literacy scores' first plausible value was used to compute the range of scores within the responding sample and to predict the range of estimates for nonrespondents.

If the weighting classes were well defined, that is, each weighting class successfully contains a homogeneous population in terms of proficiency scores, then scores would not vary much within a weighting cell, so the range of bias would be small. On the other hand, the range of bias is also affected by the response rate. If the response rate is high, the range of bias may not be high even when the respondents have a wide range of scores in the weighting cell, because the proportion of nonrespondents whose score will get filled in with the extreme values is low. Thus, the range of bias analysis measures the impact of response rate on the quality of final estimates as well as the effectiveness of the weighting adjustments in reducing the potential for bias.

Figure 16-3 displays the range of potential bias in outcome statistics after weighting adjustments are incorporated in the official weights. For comparison purposes, the range of bias before weighting is included in the figure also. The range of bias before weighting was computed without regard to weighting cells, based on the extreme assumption that nonrespondents would all score at the 10th percentile, and at the other extreme they would all score at the 90th percentile. The countries are sorted by their nonresponse rate and each country's nonresponse rate is shown in Figure 16-4. Figure 16-3 shows that the range of bias after weighting adjustment is significantly lower than before weighting adjustments are conducted, that is, each country data achieved a substantial bias reduction through nonresponse adjustment weighting. In addition, countries with higher response rates, such as Ireland, France, Slovak Republic, and Estonia, have lower range of bias. However, some countries with a low response rate, such as Sweden, Spain, Denmark, and Japan, have low ranges of bias also, due to their effective nonresponse adjustment weighting processes. Results of from Russian Federation were inconclusive due to non-compliance. The results from the range of bias analysis re-emphasizes the importance of minimizing bias in the sample throughout the survey process, and achieving high response rates

especially if the country does not have access to auxiliary variables highly correlated with proficiency.

Figure 16-3. Range of potential bias before and after weighting adjustment

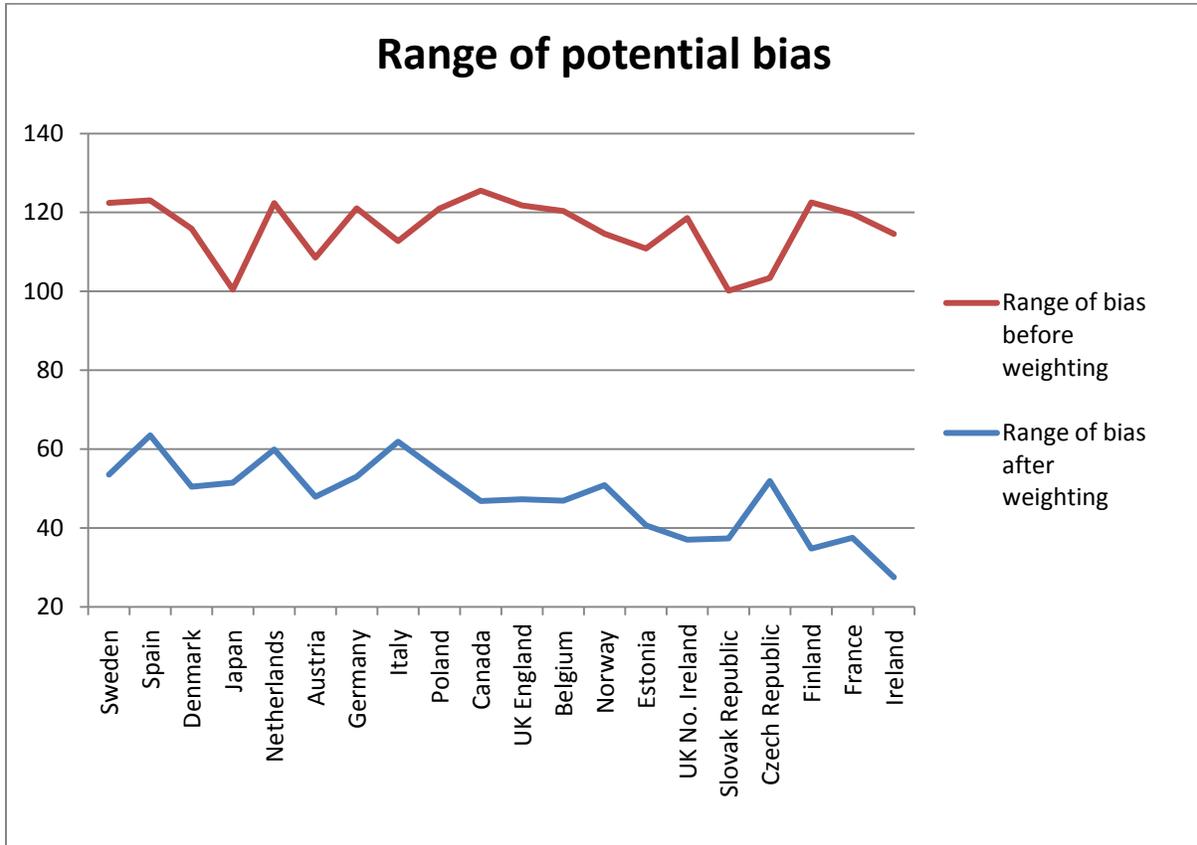
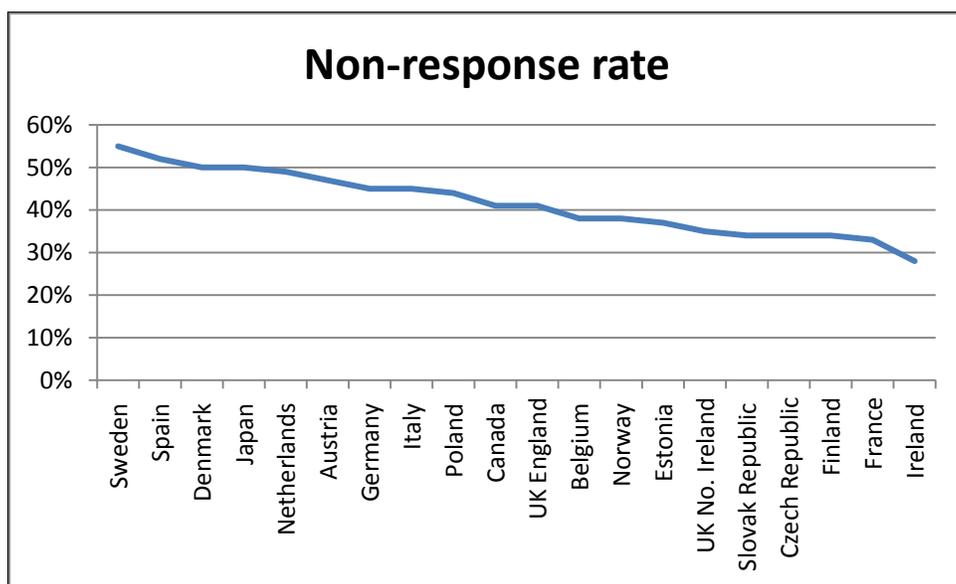


Figure 16-4. Nonresponse rate by participating country



16.3.3 Item NRBA

Countries were required to conduct nonresponse bias analysis for any BQ item with a response rate below 85%. Only two items showed low response rates: item D_Q17B (Earnings – additional payment amount last year), and item D_Q18A (Earnings – total earning last year).

Czech Republic, Estonia, Italy, Poland, Russian Federation and Slovak Republic were the only countries that had less than an 85% response rate for either D_Q17B or D_Q18A, with the lowest response rate being equal to 75% for D_Q18A for Poland.

16.3.4 Summary and conclusions

PIAAC standards were established with the main goal of producing reliable and comparable data across participating countries. As a result, a number of standards and guidelines were developed to help countries achieve the highest response rate possible, and at the same time reduce nonresponse bias to the minimum achievable. In addition, all countries were required to conduct a basic NRBA, and countries with lower response rates were required to conduct an extended NRBA.

All countries were required to conduct a basic nonresponse bias analysis (NRBA) and report the results. In addition, countries were required to conduct and report the results of a more extensive NRBA if the overall response rate was below 70%, or if any stage of data collection (screener, background questionnaire, or the assessment) response rate was below 80%. An item NRBA was required for any BQ item with response rate below 85%.

The basic and extended NRBA included several analyses. Each analysis was based a number of assumptions about nonrespondents, limiting the utility of the results. Thus, multiple analyses were used to assess the potential for bias in outcome statistics.

Correlation between the auxiliary variables used during weighting and the proficiency scores is a good indication of the effectiveness of nonresponse adjustment weighting. A number of countries with low response rates had higher correlations, implying a more effective nonresponse adjustment than countries with lower correlations. However, data users need to be cautioned that the analysis is based on correlations between respondents' proficiency scores and the auxiliary variables. That is, the analysis assumes that the same correlations exist for the remaining sampled cases that have no scores.

Table 16-6 summarizes the results of the NRBA for countries with response rates lower than 70%. The analysis showed that nonresponse adjustment weighting was effective in reducing the potential for bias in all countries. Countries that achieved higher response rates guaranteed a minimized level of bias in outcome statistics, whereas countries with lower response rates had to rely on the auxiliary variables available to them for nonresponse adjustment. Countries with relatively higher response rates and highly effective nonresponse adjustment showed minimal potential for bias as compared to countries with lower response rates, or those with moderately effective nonresponse adjustment weighting.

The analysis concluded that there was not enough evidence showing any moderate or high level of bias in the outcome statistics across the countries. However, this conclusion was based on assumptions made about the proficiency scores of nonrespondents. Therefore, data users need to be cautioned when interpreting the results of the NRBA for countries with very low response rates since different assumptions could lead to different results. For example, a response rate of 50% would mean making assumptions about half of the sample with no data. Multiple analyses, with different assumptions, were included in the NRBA to protect against misleading results, however, the lower the response rate, the higher is the risk of hidden biases that are undetectable through NRBA even when multiple analyses are involved.

Table 16-6: PIAAC NRBA outcome summary for countries with response rates lower than 70%

Country	Outcome
Austria	Caution-Bias low
Canada	Caution-Bias minimal
Czech Republic	Caution-Bias low
Denmark	Caution-Bias low
England (UK)	Caution-Bias low
Estonia	Caution-Bias low
Finland	Caution-Bias minimal
Flanders (Belgium)	Caution-Bias low
France	Caution-Bias minimal
Germany	Caution-Bias low
Italy	Caution-Bias low
Japan	Caution-Bias low
N. Ireland (UK)	Caution-Bias low
Netherlands	Caution-Bias low
Norway	Caution-Bias low
Poland	Caution-Bias low

Table 16-6 (cont.): PIAAC NRBA outcome summary for countries with response rates lower than 70%

Country	Outcome
Russian Federation ¹⁴	Caution-Bias level unknown ¹
Slovak Republic	Caution-Bias low
Spain	Caution-Bias low
Sweden	Caution-Bias low

¹ Bias level unknown due to incomplete nonresponse bias analyses.

16.4 Sample sizes and design effects

A high-quality survey produces estimates that are both unbiased and low in variability. The bias aspect was discussed in previous sections. This section will address the variability aspect. Sample size is one of the main factors that affect the variability of survey estimates. The smaller the sample size, the higher the variability of survey estimates. However, given the same sample size, the survey estimates from a simple random sample often have lower variability than those from complex sample designs. The effect of the sampling design on the variability of estimates is usually referred to as the design effect. In the following, we discuss the PIAAC sample sizes and design effects in turn.

16.4.1 Sample sizes

Table 16-7 shows the actual sample size for each country. By “actual sample size”, we refer to the number of cases with a final weight for analysis. The sample size includes both BQ respondents and BQ literacy-related nonrespondents (LRNR) with age and gender collected. The number of BQ LRNR cases is shown in a separate column as well. The BQ LRNR cases are different from the other nonrespondents because they did not complete the BQ due to literacy-related reasons, which means their proficiency levels cannot be represented by those of respondents. Therefore the percentage of such cases will be reported in data analysis although they do not have proficiency scores available.

Table 16-7: Actual sample sizes, by country

Country	Actual sample size [*]	BQ LRNR with age and gender collected
Australia	8,600	154
Austria	5,130	105
Canada	27,285	231
Cyprus ¹⁵	5,053	661
Czech Republic	6,102	21
Denmark	7,328	42
England (UK)	5,131	51
Estonia	7,632	46
Finland	5,464	0
Flanders (Belgium)	5,463	480

¹⁴ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

¹⁵ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

Table 16-7 (cont.): Actual sample sizes, by country

Country	Actual sample size*	BQ LRNR with age and gender collected
France	6,993	86
Germany	5,465	86
Ireland	5,983	20
Italy	4,621	32
Japan	5,278	105
Korea	6,667	16
Netherlands	5,170	87
Northern Ireland (UK)	3,761	35
Norway	5,128	181
Poland	9,366	0
Russian Federation ¹⁶	3,892	0
Slovak Republic	5,723	22
Spain	6,055	85
Sweden	4,469	0
United States	5,010	112

*The actual sample size is affected by several factors including response rates, number of languages, oversampling of subgroups, and the inclusion of reading components. Please refer to Chapter 14 for details.

16.4.2 Variability in sampling weights

A key component of the design effect is due to differential sampling weights. As mentioned in Chapter 14, several PIAAC countries sampled certain subgroups of population at a higher rate to obtain sufficient precision for analysis of the subgroups. For countries with a household sampling stage, people from different household sizes were also sampled with different probability. This led to unequal sampling weights and an increase in the variability of survey estimates. In addition, sampling weights were adjusted to account for sample nonresponse and undercoverage, which normally made the weights more variable. The variability of weights can be expressed by the coefficient of variation (CV) of the weights. The CV is

$$CV_w = \frac{\sigma_w}{\bar{w}},$$

where σ_w is the standard deviation of the weights and \bar{w} is the mean of weights.

Table 16-8 shows the CV of both the base weights and final sampling weights for each country. The base weights are computed as the inverse of the probability of selection, while the final weights result from the weighting adjustments.

¹⁶ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

Table 16-8: Variability in sampling weights

Country	Sample Design	CV of household base weight ¹	CV of person base weight ²	CV of person final weight
Australia	Screener	Not available ³	Not available ³	0.78
Austria	Registry	NA	0	0.30
Canada	Screener	1.31	1.28	1.33
Cyprus ¹⁷	Screener	0.03	0.51	0.63
Czech Republic	Screener	1.52	1.71	1.37
Denmark	Registry	NA	0.46	0.52
England (UK)	Screener	0.30	0.57	0.59
Estonia	Registry	NA	0	0.21
Finland	Registry	NA	0.04	0.21
Flanders (Belgium)	Registry	NA	0	0.21
France	Registry	NA	0.10	0.23
Germany	Registry	NA	0.47	0.47
Ireland	Screener	0.37	0.62	0.61
Italy	Screener	0.12	0.50	0.66
Japan	Registry	NA	0.02	0.32
Korea	Screener	0.52	0.42	0.43
Netherlands	Registry	NA	0	0.31
Northern Ireland (UK)	Screener	0.82	2.29	0.73
Norway	Registry	NA	0	0.22
Poland	Registry	NA	0.91	0.97
Russian Federation ¹⁸	Screener	0.57	1.44	1.04
Slovak Republic	Registry	NA	0	0.47
Spain	Registry	NA	0.33	0.46
Sweden	Registry	NA	0	0.36
United States	Screener	0.00	0.36	0.52

¹ Household base weights are not applicable (NA) to registry countries.

² For screener countries, the CV of person base weight is based on the person base weight described in section 15.1.3, which has the screener weighting adjustments in it.

³ Australia did not provide information on the CVs of household and person base weight because of confidentiality restrictions.

The CV of the base weights is generally larger for countries with a household sampling stage (referred to as screener hereafter) than those without a household sampling stage (referred to as registry hereafter) due to differential probabilities of selection caused by differential household sizes. Among screener countries, the United Kingdom has the largest CV of base weights due to subsampling of multiple households at the same selected addresses in Northern Ireland (UK), and the Czech Republic's CV is high due to a supplemental sample of certain age groups. Among the registry countries, Poland has the largest CV caused by oversampling of certain age groups.

¹⁷ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

¹⁸ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.

16.4.3 Design effects and effective sample sizes

Many of the PIAAC countries used complex sample designs that involved clustered samples to meet cost limitations and be operationally feasible. For example, a sample may consist of 500 street blocks (clusters) with 10 people from each block. Because people who live in the same blocks tend to have more similar social and economic background than others, a simple random sample of 5,000 people is thus likely to cover the diversity of the population better than a sample of 500 blocks with 10 people from each block. Thus, the uncertainty (i.e. standard error) associated with any population parameter estimate will be larger for a clustered sample than for a simple random sample of the same size.

Furthermore, as mentioned in the previous section, unequal sampling weights also increased the variability of survey estimates.

The design effect is expressed by the ratio of the variance of the estimate obtained from the (more complex) sample to the variance of the estimate that would be obtained from a simple random sample with the same number of sampling units. Design effects can be used to evaluate the efficiency of the PIAAC sample designs. In addition, the design effects from this study can be used to estimate initial sample sizes for the next cycle of PIAAC.

As mentioned earlier in Chapter 15, the PIAAC variance can be estimated by using the replication technique¹⁹, which accounts for the complex design (sampling and imputation error variance components as described in section 15), and a design effect can be computed for a statistic t using

$$Deff(t) = \frac{Var_{Complex}(t)}{Var_{SRS}(t)}$$

where $Var_{Complex}(t)$ is the variance for the complex sample for the statistic t computed by the replication method, and $Var_{SRS}(t)$ is the sampling variance for the same statistic on the same data but considering the sample as a simple random sample. The simple random sampling variance is computed as the average of the simple random sampling variance for each of the 10 plausible values.

Another way to express the reduction of precision due to the complex sample design is the effective sample size, which is the simple random sample size that would give the same sampling variance as the one obtained from the actual complex sample design. The effective sample size for a statistic t is

$$Effn(t) = \frac{n}{Deff(t)},$$

where n is the actual sample size.

The estimated design effects and effective sample sizes for proficiency scores over for each country are shown in Table 16-9 below.

¹⁹ The Taylor Series linearization approach can be used to estimate the numerator as well.

Table 16-9: Design effects and effective sample sizes for proficiency score, by country

Country	Design effect			Effective sample size (Literacy) ¹
	Literacy	Numeracy	Problem solving	
Australia	2.39	2.06	2.81	3,061
Austria	1.41	1.61	1.44	3,561
Canada	3.45	4.39	4.80	7,848
Cyprus ²⁰	1.54	1.25	--	2,855
Czech Republic	3.53	2.75	2.87	1,725
Denmark	1.24	1.47	1.56	5,861
England (UK)	2.33	2.03	2.18	2,176
Estonia	2.00	1.02	2.95	3,785
Finland	0.94	1.00	1.73	5,464
Flanders (Belgium)	1.55	1.34	1.45	3,215
France	1.01	0.81	--	6,867
Germany	2.01	1.89	2.58	2,680
Ireland	2.25	2.16	2.57	2,652
Italy	2.75	2.08	--	1,666
Japan	1.54	1.48	2.38	3,362
Korea	1.31	1.52	2.02	5,086
Netherlands	1.10	0.99	1.50	4,635
Northern Ireland (UK)	6.62	4.71	7.14	563
Norway	0.83	1.05	0.88	4,947
Poland	1.48	2.47	4.54	6,320
Russian Federation ²¹	15.77	16.62	22.33	247
Slovak Republic	1.35	1.58	1.74	4,236
Spain	1.27	0.88	--	4,710
Sweden	0.80	0.99	0.86	4,469
United States	2.21	2.05	2.84	2,211

¹ The effective sample size was computed as the number of cases with plausible values divided by the overall design effect. The effective sample size is set equal to the actual number of cases with plausible values for countries where the overall design effect is less than or equal to 1.

²⁰ Please refer to notes A and B regarding Cyprus in the *Note to Readers* section of this report.

²¹ Please refer to the note regarding the Russian Federation in the *Note to Readers* section of this report.