Treatment of Outcome-Related Nonresponse in an International Literacy Survey

Wendy Van de Kerckhove[1], Leyla Mohadjer[1], Tom Krenzke[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract
This paper describes efforts to address nonignorable nonresponse in the Program for the International Assessment of Adult Competencies (PIAAC), an international survey of adult literacy. A source of nonignorable nonresponse in PIAAC is from persons who cannot complete the survey because of a language barrier, reading/writing barrier, or learning/mental disability. Such persons are part of the target population. However, they cannot be represented by respondents because their reason for not completing the survey is directly related to the survey outcome (proficiency in the assessment language). We report on data collection, weighting, and estimation procedures implemented in PIAAC to limit this source of bias and allow for comparability between countries.

Key Words: Nonignorable, missing not at random (MNAR), bias, PIAAC

1. Introduction

Nonresponse can be a source of bias in survey estimates if the characteristics of the survey respondents differ from those of the nonrespondents. Weighting and imputation procedures are often used to adjust for nonresponse and reduce potential nonresponse bias. Standard procedures, such as those described in Kalton and Flores-Cervantes (2003) or Kalton and Kasprzyk (1986), are based on a missing at random (MAR) assumption. MAR means that the probability of nonresponse is independent of the survey outcome (Y) after controlling for auxiliary characteristics (X).

Nonignorable nonresponse occurs when the probability of nonresponse is related to Y, even after controlling for X. In other words, the reason for nonresponse is directly related to the survey outcome. Nonignorable nonresponse is also known as missing not at random (MNAR). An example would be persons who cannot complete a health survey because they are too ill. Work is currently being done to develop procedures to address this type of nonresponse. For example, Kott and Chang (2010) describe the use of calibration weighting, and Siddique and Belin (2008) propose multiple imputation with approximate Bayesian bootstrap.

This paper is on nonignorable nonresponse in the Programme for the International Assessment of Adult Competencies (PIAAC). PIAAC is an adult literacy survey sponsored by the Organization for Economic Cooperation and Development. A source of nonignorable nonresponse in PIAAC is from persons unable to complete the survey because of a language barrier, reading/writing barrier, or learning/mental disability. Such

reasons are directly related to the survey outcome (proficiency in the assessment language).

Section 2 will provide some background on the PIAAC survey. Section 3 will cover the approaches implemented in PIAAC to address literacy-related nonresponse (LRNR). It will describe efforts to limit and monitor LRNR during data collection, the use of separate nonresponse adjustments for LRNR and other nonrespondents during weighting, and the creation of two sets of estimates for analysis – one excluding LRNR and one using minimum value imputation. Section 4 will include a brief discussion on plans to improve this process in future cycles of the survey. At the time of this paper, the results of the survey have not yet been released to the public, so the focus is on methodology rather than specific outcomes.

2. PIAAC Survey Design

The purpose of PIAAC is to assess the level and distribution of adult skills across countries, focusing on the cognitive and workplace skills needed for successful participation in the economy and society of the 21st-century. Twenty-four countries participated in Round 1 of PIAAC, with data collection ending in 2012. At the time of this paper, an additional nine countries are taking part in the field test for Round 2. Countries are responsible for sample selection and data collection, and have the option of implementing the weighting procedures. The PIAAC Consortium is responsible for weighting and estimation, as well as providing standards and implementing quality checks for the entire survey process.

2.1 Sample Design and Data Collection

The PIAAC target population consists of non-institutionalized adults aged 16 to 65. Adults are to be included regardless of citizenship, nationality, or language. The core sample design is a multi-stage area sample; however, this differs depending on the available sampling frame and operational constraints within the country. Some countries are able to sample persons directly from a population registry. Other countries have no such registry and must sample dwelling units and then conduct a screener to enumerate and sample persons within the dwellings.

The survey is conducted face-to-face using computer assisted person interviewing (CAPI). It involves two or three stages of data collection: a screener (if needed), a background questionnaire (BQ), and an assessment. The BQ collects information on demographic characteristics, educational and employment experiences, and literacy-related activities. The assessment includes a series of literacy, numeracy, and problem solving tasks.

The survey is intended to measure proficiency in the language of the assessment. Table 1 lists the 24 countries that participated in Round 1 of PIAAC, along with the language or languages of the assessment. For countries that provided it, it also indicates the percentage of the population that speaks the assessment language, as reported by the countries. For some countries, close to 100 percent of the population speaks the assessment language(s), while for others five percent or more of the population speaks a language other than the assessment language(s). This provides an indication that there may be varying levels of nonignorable nonresponse among countries.

Table 1: Assessment Languages for PIAAC Round 1 Countries

Country	*Assessment language(s) and proportion of population speaking it (as available)*
Australia (AUS)	English
Austria (AUT)	German (88.5%)
Belgium-Flanders (BEL)	Dutch
Canada (CAN)	Canadian English (67.3%), French (21.1%)
Cyprus (CYP)	Greek (84.1%)
Czech Republic (CZE)	Czech
Denmark (DNK)	Danish (92%)
Estonia (EST)	Estonian (67%), Russian (33%)
Finland (FIN)	Finnish (90.5%), Swedish (5%)
France (FRA)	French
Germany (DEU)	German
Ireland (IRL)	English
Italy (ITA)	Italian
Japan (JPN)	Japanese (~100%)
Korea (KOR)	Korean
Netherlands (NLD)	Dutch
Norway (NOR)	Norwegian (Bokmål)
Poland (POL)	Polish
Russian Federation (RUS)	Russian (98.2%)
Slovak Republic (SVK)	Slovak (89.8%), Hungarian (10.2%)
Spain (ESP)	Castellano (60%), Gallego (6%), Catalan (18%), Valencian (11%), Euskera (5%)
Sweden (SWE)	Swedish
United Kingdom (GBR)	UK English
United States (USA)	English (91.5%)

2.2 Weighting and Estimation

The final proficiency estimates are produced using weights that reflect the sample design and are adjusted to reduce differences between the respondents and target population. First base weights are created as the inverse of the selection probabilities. Then the weights are adjusted for unknown eligibility[1] and nonresponse to the screener and BQ. The resulting weights are calibrated to control totals (if necessary, extreme weights were trimmed and recalibrated) to produce the final weights for analysis.

The final outcome measure is proficiency scores, ranging from 0 to 500. Sampled persons who complete the BQ but not the assessment receive a final weight and have their proficiency scores imputed based on the BQ data. In addition, the survey is designed such that not every respondent receives the same set of assessment items. Therefore, Item Response Theory (IRT) modeling is used to impute 10 plausible values to each respondent.

[1] For countries with a screener stage, the interviewer might not be able to determine whether a household contains anyone between the ages of 16 and 65. Similarly, for countries with a registry, it may be unknown whether the sampled person is eligible because of an inability to locate the individual. A portion of the weights of the cases with unknown eligibility status is distributed to the ineligible cases.

3. Addressing Nonignorable Nonresponse

PIAAC countries are instructed on the treatment of LRNR through the Technical Standards and Guidelines and an International Interviewer Procedures Manual. These instructions are intended to standardize the treatment of such cases across countries and ensure consistent estimates. They cover procedures for addressing LRNR during data collection, weighting, and estimation.

3.1 Data Collection

During data collection, emphasis is placed on minimizing the amount of LRNR. Countries are encouraged to translate the BQ (and screener) into multiple languages and to utilize bilingual interviewers. If no bilingual interviewer is available in the area, a family member or neighbor can serve as an interpreter for the BQ.

In addition, it is important to identify and capture LRNR. Disposition codes have been developed for recording the reason for nonresponse. These include codes for language barriers, reading/writing barriers, and learning/mental disabilities. Countries are also asked to collect age and gender of the LRNR cases. This is done so that at least minimal information is available on these cases for domain analysis and potential use in modeling.

Countries are required to continuously monitor the levels of LRNR during data collection and report the count of such cases each month. In addition, they are asked to review the characteristics of these cases, such as their age, gender, and the region in which they reside. This provides a check on whether the disposition codes are being used correctly. It can also indicate whether any additional efforts are needed to reduce the amount of LRNR.

Figures 1 and 2 show the results of the data collection strategies for Round 1. They indicate the weighted percentage of LRNR for each country by data collection stage, where Figure 1 includes countries with two stages of data collection and Figure 2 is for those with three stages. Both plots indicate that the majority of LRNR occurred at the BQ stage. In addition, the prevalence was generally around two percent or less for any stage of data collection.

However, for two countries, the amount of LRNR to the BQ exceeded this level. Both countries offered the BQ in only one language (although the screener was offered in multiple languages for the country in Figure 1). Approximately 90% of the BQ LRNR for these countries was due to language barriers.

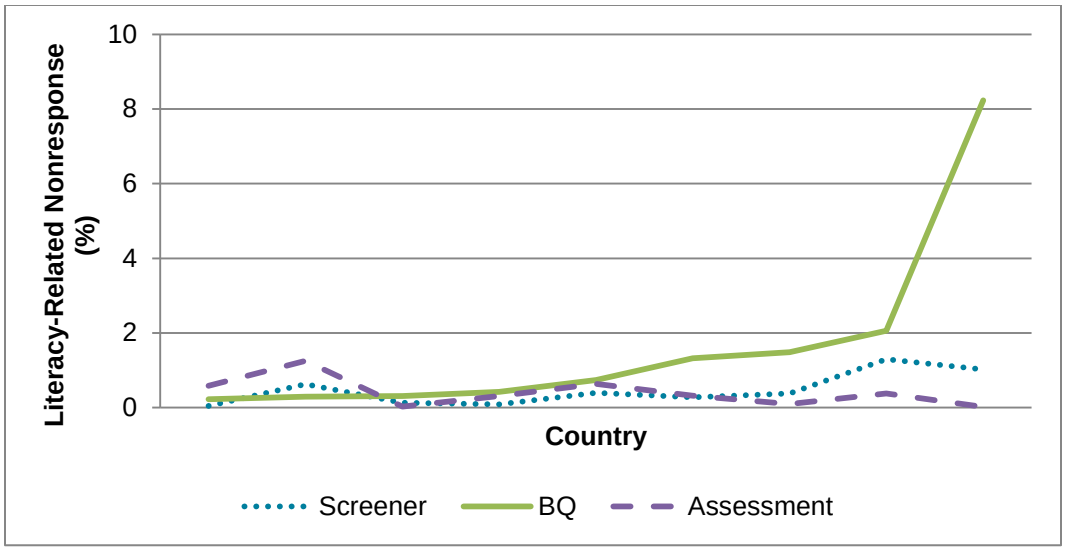

Figure 1: Weighted percentage of literacy-related nonresponse among sampled cases, screener countries

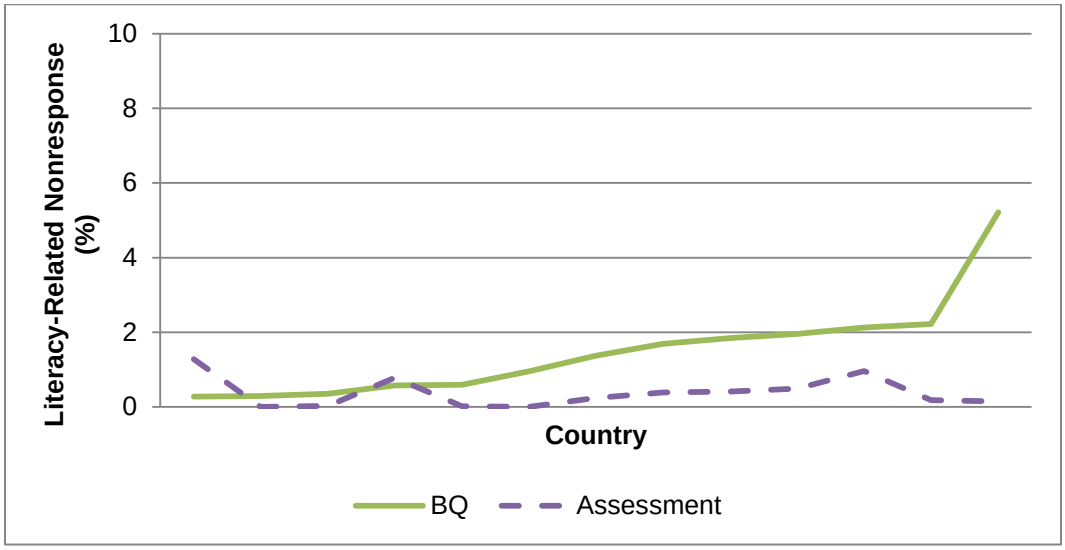

Figure 2: Weighted percentage of literacy-related nonresponse among sampled cases, registry countries

3.2 Weighting

As noted in section 1, standard nonresponse weighting adjustments have a MAR assumption. Tables 2 and 3 provide a simple illustration of the impact of treating nonresponse as MAR for PIAAC. The tables are identical except Table 2 has two percent of sampled cases as LRNR, whereas in Table 3 it is eight percent. In this example, the sample is divided into two subgroups – A and B – which serve as the two nonresponse weighting cells. For example, this could be the East and West region of the country. The response rate for subgroup A is 62.5%, and the mean score of respondents is 300. In subgroup B, the response rate is 50%, and the mean score of respondents is 250.

Standard nonresponse adjustments assume that the mean score of nonrespondents is the same as that of respondents within the same subgroup. Under this assumption, the overall

mean score is 275, as shown in column $\bar{y}_1$ of the tables. However, this assumption is not expected to hold for PIAAC. Generally, the LRNR would be expected to score at lower proficiency levels than respondents, even within the same nonresponse adjustment subgroup. Columns $\bar{y}_2$ to $\bar{y}_5$ show the change in the overall score if the assumption for LRNR changes. While there is little impact on the overall score when LRNR is at 2%, the difference is more substantial when it is at 8%. At higher levels of LRNR, the standard weighting procedures could introduce significant bias. For example, in Table 2, if the average score among 80 LRNR cases was 100, the overall mean would be 271.5. But, as shown in Table 3, if the average score among 320 LRNR cases was 100, then the overall mean falls to 261.0, which is far lower due to many more LRNR cases. The differences seen in Table 3 could have a meaningful impact on a country's ranking.

Table 2: Impact of Treating Nonresponse as MAR (2% LRNR)

Adjustment cell	*Response status*	*Weighted sample size*	*Mean proficiency score*				
			$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$
A	Respondent	2500	300	300	300	300	300
	Nonrespondent	1420	300	300	300	300	300
	LRNR	80	**300**	**250**	**200**	**150**	**100**
	Overall	4000	300	299	298	297	296
B	Respondent	2000	250	250	250	250	250
	Nonrespondent	1920	250	250	250	250	250
	LRNR	80	**250**	**250**	**200**	**150**	**100**
	Overall	4000	250	250	249	248	247
Overall	Overall	8000	**275**	**274.5**	**273.5**	**272.5**	**271.5**

Table 3: Impact of Treating Nonresponse as MAR (8% LRNR)

Adjustment cell	*Response status*	*Weighted sample size*	*Mean proficiency score*				
			$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$
A	Respondent	2500	300	300	300	300	300
	Nonrespondent	1180	300	300	300	300	300
	LRNR	320	**300**	**250**	**200**	**150**	**100**
	Overall	4000	300	296	292	288	284
B	Respondent	2000	250	250	250	250	250
	Nonrespondent	1680	250	250	250	250	250
	LRNR	320	**250**	**250**	**200**	**150**	**100**
	Overall	4000	250	250	246	242	238
Overall	Overall	8000	**275**	**273**	**269**	**265**	**261**

For this reason, the PIAAC standards state that persons who do not complete the survey for a literacy-related reason cannot be represented by survey respondents. Therefore, an alternative weighting procedure is needed. Since countries have the option of letting the Consortium create the weights or doing so themselves, this procedure must be straightforward enough to be implemented consistently across countries.

This is accomplished by having separate nonresponse adjustments for LRNR and all other nonresponse. As was shown in Figures 1 and 2, there are insufficient LRNR at the assessment stage to represent those at previous stages, so both the BQ LRNR (with age and gender collected) and the assessment LRNR receive final weights.

Specifically, for countries with a screener stage[2], the screener respondents are first weighted up to account for the non-literacy-related nonrespondents to the screener, using the following adjustment factor:

$$F_i^{SCRNR} = \begin{cases} 1 \; if \; i \in L^{SCR} \\ \frac{S'_{RSCR}+S'_{NRSCR}}{S'_{RSCR}} \; if \; i \in R^{SCR}, \\ 0 \; if \; i \in NR^{SCR} \end{cases}$$

where

S': Sum of household base weights[3] (W_i) within the same adjustment cell as household *i*
L^{SCR}: LRNR to the screener
R^{SCR}: Screener respondent
NR^{SCR}: Non-literacy-related nonrespondent to the screener

The resulting weight is $W_i F_i^{SCRNR}$.

Then the weights of the BQ respondents are adjusted by the following factor to account for the non-literacy-related nonrespondents to the BQ:

$$F_{ij}^{BQNR} = \begin{cases} 1 \; if \; j \in L^{BQA} \\ \frac{S_{RBQ}+S_{NRBQ}}{S_{RBQ}} \; if \; j \in R^{BQ}, \\ 0 \; if \; j \in NR^{BQ} \end{cases}$$

where

S: Sum of person base weights within the same adjustment cell as person *j* in household *i*, where the person base weight (W_{ij}) = $W_i * F_i^{SCRNR} *$ (1/within-household selection probability)
L^{BQA}: LRNR to the BQ or assessment
R^{BQ}: BQ respondent (excluding LRNR to the assessment)
NR^{BQ}: Non-literacy-related nonrespondent to the BQ

The resulting weight is $W_{ij} F_{ij}^{BQNR}$.

Finally, the weights of the BQ and assessment LRNR are adjusted to account for those at the screener stage, using the following adjustment factor:

$$F_{ij}^{LR} = \begin{cases} 1 \; if \; j \in R^{BQ}, \; NR^{BQ} \\ \frac{S'_{LSCR} + S'_{LBQA}}{S'_{LBQA}} \; if \; j \in L^{BQA} \\ 0 \; if \; j \in L^{SCR} \end{cases}$$

The resulting weight is $W_{ij} F_{ij}^{BQNR} F_{ij}^{LR}$.

[2] The process for registry countries is similar, with assessment LRNR and BQ LRNR with age and gender collected representing LRNR without age and gender collected.
[3] After adjusting for unknown eligibility

The nonresponse adjusted weights are trimmed, if necessary, and calibrated to population totals. This results in final weights for the following groups:

1. Assessment respondents
2. BQ respondents who are non-literacy-related nonrespondents to the assessment
3. BQ respondents who are LRNR to the assessment
4. LRNR to the BQ

3.3 Estimation

After creating final weights, proficiency scores are generated for cases in each of the four groups above. Persons who did not respond to the assessment for a reason that is not related to literacy (group 2) have their proficiency scores imputed using the BQ data. This is done based on the assumption that they are similar to the assessment respondents with the same BQ characteristics. However, this assumption cannot be made for the LRNR (groups 3 and 4).

To impute proficiency scores to LRNR to the BQ, one option would be to assume such persons would score at the lowest level of literacy. This could be accomplished by imputing wrong answers to the assessment. This was the procedure that was applied to assessment LRNR in the 2003 Survey of Adult Literacy and Life-skills and the 1994 International Adult Literacy Survey (Murray, Kirsch and Jenkins 1998). However, in PIAAC, not all countries agree with this assumption and have argued that many of the LRNR in their country are proficient in the assessment language but it is just not their preferred language.

Another complication in assigning proficiency scores is that little information is known about the BQ LRNR, since they do not have the BQ data. Some countries have additional information on such cases from their population registries, but others only know the age and gender. Age and gender was found to be insufficient for modeling.

To address these concerns, two estimation approaches are being taken for handling LRNR. The first is to report the percentage of LRNR along with the mean score for the rest of the population. When analyzing results, both measures should be considered together, making it somewhat more difficult to rank countries. The second method is to use minimum value imputation, assigning the LRNR a low proficiency score. This may be an underestimate of the actual proficiency. However, the results with and without the imputation can provide lower and upper bounds.

4. Discussion

Efforts are being made at all stages of the PIAAC survey to address nonignorable nonresponse resulting from persons who cannot complete the survey because of a language barrier, reading/writing barrier, or learning/mental disability. During data collection, countries are encouraged to use translations and interpreters for the BQ and are required to capture and monitor reasons for nonresponse. To produce final weights, separate nonresponse adjustments are implemented for LRNR and other nonrespondents so that the LRNR will not be represented by respondents. The LRNR are also considered separately in estimation.

Experience from Round 1 indicates that for the majority of countries, LRNR occurs at low levels and should have minimal impact on the resulting estimates. However, this does not hold for all countries. If not handled appropriately, LRNR could introduce significant bias in the PIAAC estimates and have a meaningful impact on a country's ranking. In addition, it was found that most of the LRNR occurs at the BQ stage and is due to language barriers.

Therefore, the focus in future cycles of PIAAC will be to minimize the amount of LRNR to the BQ. The current plan is for countries to offer the BQ into as many languages as possible. To reduce the costs associated with translation, countries can borrow BQ's from other countries. If this is successful, it could address some of the complications associated with estimation, as described in section 3.3, and the estimation procedures can be re-evaluated. We also recommend continuing the special weighting procedures for LRNR.

References

Kalton, G. and I. Flores-Cervantes. 2003. Weighting methods. In *Journal of Official Statistics*. 19(2), 81-97.

Kalton, G. and D. Kasprzyk. 1986. The treatment of missing survey data. In *Survey Methodology*. 12, 1-16.

Kott, P.S. and T. Chang. 2010. Using calibration weighting to adjust for nonignorable unit nonresponse. In *Journal of the American Statistical Association*. 105(491), 1265-1275.

Murray, T.S., I. Kirsch, and L. Jenkins. 1998. Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey. NCES 98-053, U.S. Department of Education, Office of Educational Research and Improvement, Washington, DC.

Siddique, J. and T.R. Belin. 2008. Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. In *Computational Statistics & Data Analysis*. 53(2), 405-415.