Salvaging Data from an Incomplete Sample Through Statistical Data Integration

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

Abstract
Incomplete survey data can arise when there are unexpected disruptions to data collection, such as the recent pandemic. The result is a sample that is a product of the probability-based sample design, the non-probabilistic mechanism that determined which sampled cases were worked prior to the disruption, and nonresponse. We describe a method used in the U.S. Program for the International Assessment of Adult Competencies (PIAAC) for combining incomplete survey data with complete survey data. The sample design consisted of a nationally representative core sample and a supplemental state-based sample, where the goal of the supplemental sample was to improve small area estimates. Given the uncertainty in response rates and potential pandemic-related disruptions, the design allowed for the flexibility to shift fieldwork effort from the supplemental sample to the core sample as needed. In fact, data collection for the state supplement was halted less than halfway into the data collection period, before interviewers had visited all areas. To salvage the collected data from the incomplete supplemental sample, we combined it with the core sample by using a composite weighting technique. We describe the weighting methods and an evaluation of the combined data.

Key Words: incomplete data, composite weighting, hybrid weighting, non-probability

1. Introduction

With declining response rates for probability samples and an increased demand for timely data, the use of administrative data, web panels, or other non-probability samples has become increasingly commonplace. Non-probability samples can also arise unexpectedly. For the U.S. Programme for the International Assessment of Adult Competencies (PIAAC) Cycle 2, the situation was one of an incomplete sample, meaning a sample that started out as a probability sample, but then there were unexpected disruptions so data collection could not be completed. Other organizations may have experienced a similar situation during the pandemic, when in-person surveys were halted. The resulting sample is the outcome of probability selection, the non-probabilistic mechanism that determined which sampled cases and areas (primary sampling units) were worked prior to the disruption, and nonresponse.

Given the time, effort, and cost in collecting the data, a solution is needed to salvage the data and make valid inferences. If a corresponding probability sample exists, then methods for combining probability and non-probability samples can be applied to this situation. This was the case for U.S. PIAAC Cycle 2, which is described further in section 2. Many methods for combining probability and non-probability samples exist in the literature. Dever (2018) discusses "Hybrid" estimation while combining a data file containing probability-based and nonprobability sample cases. Dever explores composite estimation and propensity score adjustment as ways to have the combined data project to the intended target population. Valliant (2020) compares seven alternatives for estimation from nonprobability samples. For example, a quasi-randomization approach is applied to estimate pseudo-inclusion probabilities for the non-probability sample to correct for selection bias. The approach uses a binary regression model to help combine the non-probability sample with a probability-based reference sample to estimate pseudo-inclusion probabilities. The reference sample needs to represent the target population through its sample weights. Wu (2022) classified inference methods for non-probability samples into three types: (1) model-based prediction approach, which uses regression models or mass imputation to predict the outcome, (2) inverse probability weighting, which models propensity scores to create weights, and (3) doubly-robust estimation, which combines (1) and (2).

For our situation, we opted for a calibration and compositing approach, which builds on Dever (2018) and is described in section 3. Including the incomplete data from the supplemental sample had a negligible effect on the resulting national estimates, as shown in the evaluation in section 4, and is expected to improve model-based estimates.

2. U.S. PIAAC Cycle 2

PIAAC is an international survey of working-age adults sponsored by the Organization for Economic Cooperation and Development (OECD). It assesses skills in literacy, numeracy, and problem-solving through an in-person household survey. The survey consists of a Background Questionnaire (BQ) and a cognitive assessment. PIAAC Cycle 1 involved of three rounds of data collection between 2011 and 2018, with 38 countries participating in at least one of the three rounds. Data collection for PIAAC Cycle 2 occurred in 2022 to 2023, with 31 countries participating.

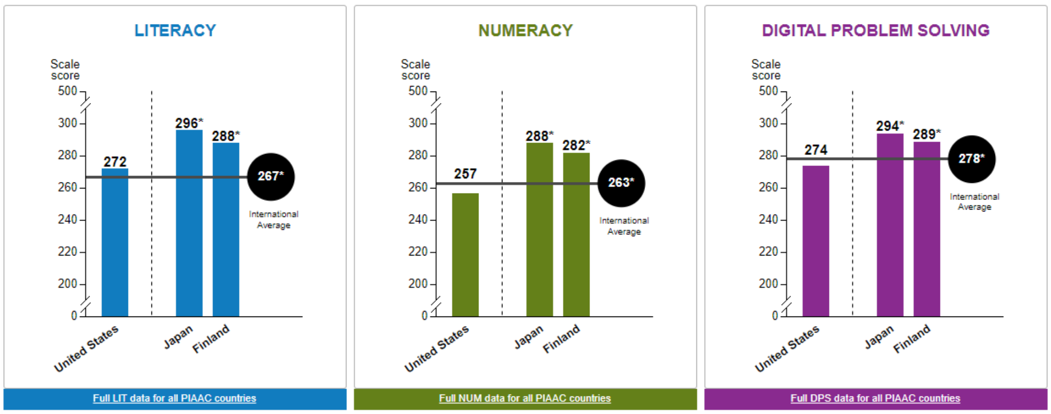
The United States collected data in all rounds to date, with sponsorship by the National Center for Education Statistics (NCES). The target population for U.S. PIAAC consisted of non-institutionalized adults aged 16 to 74. The age range of 16 to 65 is consistent with the international target population and is used for international comparisons. Adults aged 66 to 74 are of particular interest to the United States and so were added as a country-specific sample. Results from U.S. PIAAC Cycle 1 can be found at https://nces.ed.gov/surveys/piaac/. A first look at national results and international comparisons from U.S. PIAAC Cycle 2 will be available in December 2024.

2.1 Design Goals

The goal of the U.S. PIAAC Cycle 2 design was to provide high quality data and sufficient yield for national estimates and modeling purposes, specifically small area estimation (SAE) models and psychometric models. As part of an international survey, it had to adhere to international standards and guidelines (OECD 2022) as well as NCES statistical standards (NCES 2012).

U.S. PIAAC Cycle 2 was designed to produce national estimates of mean proficiency scores and proficiency-level distributions for the three proficiency domains – literacy, numeracy, and adaptive problem solving. The sample size needed to be sufficient to produce estimates for key sub-populations, such as age group, and to make comparisons with other countries in Cycle 2 and for the United States across cycles. Figure 1 shows an example of the reporting of national estimates from Cycle 1. It provides a comparison of the mean proficiency scores[1] for adult aged 16 to 65 in the United States against the two highest performing countries and the international average. NCES will publish similar reports for Cycle 2.

Figure 1-A. Average scores on PIAAC literacy, numeracy, and digital problem solving for adults age 16 to 65 for the United States and highest-performing countries: 2012–15



* Significantly different ($p < .05$) from the United States.

NOTE: LIT = Literacy. NUM = Numeracy. DPS = Digital problem solving. Average scores for the United States are compared to the PIAAC international average and highest-performing countries. Results for 21 of the countries were gathered in 2011–12, and an additional 9 participated in 2014–15. The two highest-performing countries are shown, in descending order from left to right within each domain. Results for the United States are shown on the far left within each domain to highlight that comparison.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Organization for Economic Cooperation and Development (OECD), Program for the International Assessment of Adult Competencies (PIAAC), 2012–15.

Figure 1. Example of the reporting of national estimates for U.S. PIAAC, taken from the Cycle 1 *PIAAC International Highlights Web Report* (NCES 2020-127)

Another goal of Cycle 2 is to use SAE models to provide indirect estimates of proficiency for counties, states, and sub-populations within states. A similar effort was made in Cycle 1, resulting in the PIAAC Skills Map. The skills map allows users to view proficiency estimates for counties, states, and age or education domains within states and counties; make comparisons across counties, states, or with the nation; and obtain additional demographic information for the area. It can be a useful tool for state adult education and labor departments in making policy or program decisions. Figure 2 shows an example of a heat map taken from the skills map. It indicates the percentage of the population at or below Level 1 in literacy (the lowest level) for each county in the United States, with darker shades of blue indicating a higher percentage.

To produce high quality small area estimates, the aim in Cycle 2 was to maximize the number of states with PIAAC survey data so there would be less reliance on the model to extrapolate for states without survey data. We also planned to select at least two counties per state. This would allow us to produce direct estimates of variance for input to the model. It would also better capture the diversity in proficiency between counties in the same state.

[1] Adaptive Problem Solving in Cycle 2 replaced the Cycle 1 proficiency domain of Digital Problem Solving.

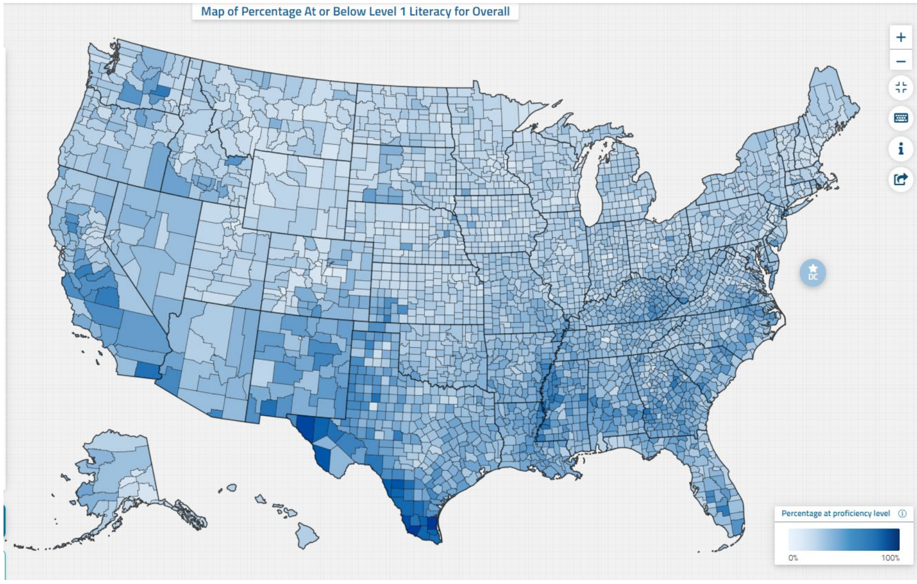


Figure 2. Example of the reporting of small area estimates for U.S. PIAAC, taken from the Cycle 1 Skills Map at https://nces.ed.gov/surveys/piaac/skillsmap/.

Finally, the international standards required a sufficient sample size for psychometric models. PIAAC involves a proficiency assessment. Given the large number of test items, not every participant receives every item. Therefore, the final proficiency scores cannot be determined by simply adding up the number of correct items. Instead, proficiency plausible values (PVs) are generated using Item Response Theory (IRT) and latent regression modeling (Yamamoto, et al., 2019). The PIAAC Consortium required a minimum number of responses per item for this purpose.

2.2 Sample Design

The U.S. PIAAC Cycle 2 sample was designed to obtain the above objectives. The design consisted of two components: a nationally representative core sample and a state-based supplemental sample. The purpose of the supplemental sample was to obtain at least two sampled primary sampling units (PSUs) in each state when combined with the core sample, strengthening the small area estimates. Given the uncertainties in data collection coming out of the pandemic, the design specifically allowed for the possibility of dropping the state-based sample.

The core sample design consisted of a four-stage area sample, with:

1. PSUs defined as counties or groups of counties,
2. Secondary sampling units (SSUs) defined as Census blocks or groups of Census blocks,
3. Dwelling units, and
4. Eligible adults.

At the first stage, we selected a stratified, probability proportionate to size (PPS) sample of 80 PSUs. The major strata were defined by Census region, metropolitan statistical area (MSA) status, and the Cycle 1 county-level small area estimates of the percentage at or below Level 1 in literacy. Within the major strata, the PSUs were further stratified by characteristics related to educational attainment, poverty status, race/ethnicity, employment status, health insurance status, marital status, and occupation. This was done via a nested stratification process, as discussed in Krenzke and Haung (2009). The core sample yielded 4,287 respondents.

After selecting the PSUs for the core sample, states were put into three categories: (1) two or more sampled PSUs in the core sample, (2) one sampled PSU in the core sample, (3) no sampled PSUs in the core sample. States that fell into the first category had no PSUs selected for the state supplement. For those falling in the second category, we selected one additional PSU in the state through conditional sampling. The core sample PSU was removed from the frame, and one additional PSU was selected with PPS. For states in the third category, we selected two PSUs in the state using stratified PPS sampling, where the two strata were defined by the Cycle 1 county-level small area estimates of the percentage at or below Level 1 in literacy. Within PSUs, we selected SSUs, dwelling units, and eligible adults following the sample design for the core sample.

Around two months into the 9.5-month field period, data collection for the state supplemental sample was halted due to funding. At that time, the supplemental sample had yielded 350 respondents, around 20 percent of the target. Because interviewer availability differed across PSUs, the sample had not been evenly worked. Interviewers had not started contacting sampled households in some PSUs. In PSUs where work had started, some sampled households or adults could have received multiple contact attempts and been fully worked to protocol, while others could have only received a single attempt or no attempts. If proficiency levels differed between the areas and cases that had been worked versus those that had not, or between the low level-of-effort cases versus the high-level-of-effort cases, then the resulting sample could produce biased proficiency estimates if not handled appropriately in weighting and estimation.

3. Method for Integrating Samples

In deciding on a methodology for integrating the core and supplemental samples, we considered several factors:

- Our end-product was a set of analysis weights. The weights are used in the psychometric models that produce the proficiency scores and are included on public-use files for analysts.
- The "selection" probabilities for the combined sample were unknown, both at the individual and PSU level. Although we knew the initial probability of selection, we did not know the probability that a case (or PSU) was attempted. Therefore, we could not create initial design weights as the inverse of the selection probabilities.
- The outcome variables (proficiency scores) were not available at the time of weighting. As noted above, the weights are used in the creation of proficiency scores. Therefore, model-based prediction approaches, such as mass imputation, and doubly robust estimation approaches were not applicable.
- The core national sample was much larger (around ten times larger) than the state supplemental sample. This is the opposite of the typical situation in which a large non-probability sample is combined with a much smaller probability sample. Therefore, bias reduction was not as much a concern, and we could choose an approach that placed more emphasis on limiting variance, specifically variation due to unequal weights.

Given the above factors, we opted for a calibration and compositing approach to produce sampling weights, with an initial weight of 1 for the supplemental sample respondents. Figure 3 provides an overview of the process. Prior to combining the samples, the core national sample was assigned initial weights that reflected the selection probabilities and were adjusted for nonresponse. The respondents from the supplemental sample were assigned an initial weight of 1. Both samples were then calibrated to population totals and composited. The calibration and compositing process is described further in section 3.1. The composited sample was then calibrated (raked) to population totals related to age, gender, region, race/ethnicity, education, and country of birth. Extreme weights were identified and trimmed, and then another round of raking was performed. Details on the weighting process can be found in the U.S. PIAAC Technical Report for Cycle 2.

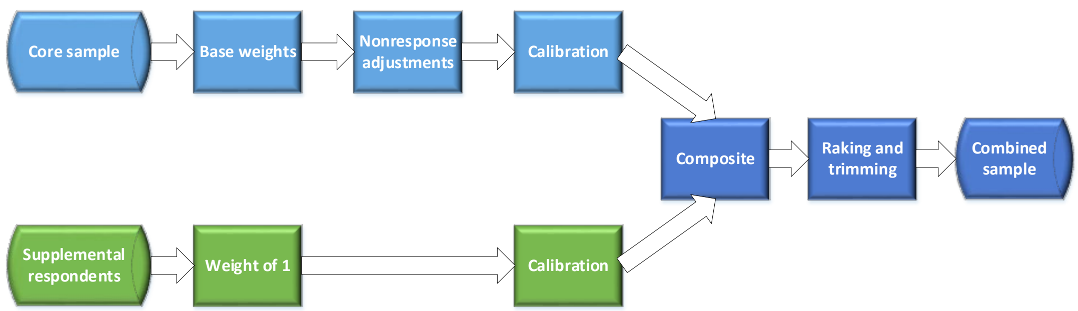


Figure 3: Weighting process for combining the PIAAC Cycle 2 core national sample and incomplete state supplemental sample

If inclusion of the supplemental sample is independent of key survey outcomes after conditioning on the calibration variables, then the chosen approach enables unbiased estimation. Although this is often a difficult assumption to meet, PIAAC has a strong set of weight calibration variables for both the pre-compositing and post-compositing calibration. By strong, we mean that the variables, such as education, have a strong correlation with proficiency (Krenzke et.al., 2019). By avoiding a more complex process, such as one involving additional nonresponse adjustments for the supplemental sample, we are also limiting variability in the weights.

3.1 Compositing

Before combining the samples, each was post-stratified to 2022 American Community Survey (ACS) control totals for age group by region by education.[2] Post-stratification cells are shown in Table 1. We could not do the full three-way crossing of the variables because of small sample sizes. In determining the post-stratification cells, we had considered using the major strata from the core sample design. However, the smallest geographic unit in the ACS data (Public Use Microdata Areas, PUMAs) crossed stratum boundaries, and control totals by major strata could not be derived. The final set of post-stratification variables met PIAAC-specific objectives and were intended to reduce inclusion bias associated with the incomplete supplemental sample. Age was chosen to distinguish between the international target population of age 16-65 and the U.S.-specific population of age 66-74. Region is related to the proportion of supplemental sample cases that were worked, as the timing of

[2] Four PSUs were selected with certainty (probability 1) in the core national sample. The four PSUs were excluded from the pre-compositing calibration and compositing process because they are self-representing and had no chance of selection for the state supplemental sample.

which sampled cases were worked varied across states according to logistical factors during the field period. Education was included as it is highly related to proficiency. By choosing a variable that was related to the inclusion probability, and a variable that was related to the survey outcome, we expect to reduce selection bias associated with the supplemental sample (Elliott and Valliant, 2017).

Table 1*:* Post-stratification cells for pre-compositing calibration

Cell	*Age group*	*Region*	*Education level*
1	16-65	1	-
2	16-65	2	-
3	16-65	3	Non-college graduates
4	16-65	3	College graduates
5	16-65	4	Non-college graduates
6	16-65	4	College graduates
7	66-74	-	Non-college graduates
8	66-74	-	College graduates

Next, composite weights were created for person *i* in domain *g* as follows:

$$\widetilde{W}_{gi}^{F} = \alpha_g^C W_{gi}^C \mathrm{I}_C(i) + \left(1 - \alpha_g^C\right) W_{gi}^S \mathrm{I}_S(i),$$

Where

C = core sample,
S = supplemental sample,
g = age 16-65 by region and age 66-74, and
α = compositing factor.

The compositing factor attempted to give more weight to the sample with lower mean-square error (MSE). It was calculated as follows, based on Krenzke and Mohadjer (2020):

$$\alpha_g^C = \frac{\frac{n_g^C}{\left(1+(cv_g^C)^2\right)\left(1+d_{g*}^C\right)}}{\frac{n_g^C}{\left(1+(cv_g^C)^2\right)\left(1+d_{g*}^C\right)} + \frac{n_g^S}{\left(1+(cv_g^S)^2\right)\left(1+d_{g*}^S\right)}},$$

Where

n_g^C and n_g^S = the number of respondents in domain *g* of the core sample and supplemental sample, respectively,
cv_g^C and cv_g^S = the coefficient of variation of the post-stratified weights for domain *g* of the core sample and supplemental sample, respectively,
d_{g*}^C and d_{g*}^S = Kolmogorov-Smirnov (K-S) statistics, described below, for group g* of the core sample and supplemental sample, respectively, and
g* = age 16-65 and age 66-74.

The term $\frac{n}{1+cv^2}$ is the effective sample size, reflecting the design effect due to unequal weights. It represents the variance component of the MSE. The term *d* represents the bias component and is the K-S statistic (Chakravart, Laha, and Roy, 1967) for the distance between the detailed (9-category) educational attainment distribution for the sample compared to the ACS. Because of limited sample size, *d* was calculated for each age group (g*) instead of by compositing domain.

Table 2 shows the resulting compositing factors and components. As indicated in the table, the core sample was considerably larger than the supplemental sample. The relative sizes varied by region, which was a reason for using region in the compositing. The design effect is larger for the core national sample because of different selection probabilities and the nonresponse adjustments. As expected, the K-S statistics are larger for the supplemental sample. The compositing factor places most of the weight on the core national sample, although there is some variation by domain. The compositing factor likely over-penalized the higher design effects for the core sample and could have been improved by putting the squared CV and the K-S statistic on a similar scale.

Table 2: Compositing Factors and Components

Domain (g)	n_g^C	n_g^S	$1 + (cv_g^C)^2$	$1 + (cv_g^S)^2$	d_{g*}^C	d_{g*}^S	α_g^C
1	463	29	1.30	1.00	0.017	0.075	0.928
2	554	34	1.27	1.00	0.017	0.075	0.931
3	1,562	110	1.35	1.02	0.017	0.075	0.919
4	711	110	1.39	1.00	0.017	0.075	0.831
5	768	67	1.29	1.03	0.036	0.084	0.906

3.2 Variance Estimation

For estimating variances, the PIAAC data tools[3] require the use of replication methods. For Cycle 2, the PIAAC Consortium recommended Balanced Repeated Replication (BRR) with a Fay's adjustment (commonly referred to as Fay's method). One reason for the recommendation was that Fay's method has been shown to be more robust in estimating variances of quantiles compared to the jackknife method (Judkins, 1990). In addition, by avoiding setting any replicate weights to zero, Fay's method avoids undefined ratio statistics for small domains (Judkins, 1990) and reduces disclosure risk (i.e., a user cannot identify a set of cases in the same cluster based on a zero replicate weight).

For the core national sample, first stage units were paired to form variance strata. The core national sample had a one-PSU-per-stratum design, so variance strata were formed by pairing non-certainty PSUs across strata. For certainty PSUs, groups of SSUs were paired to form the variance strata. Using Fay's method and a perturbation factor k of 0.3, we created 44 replicate weights.

For the supplemental sample, we created variance strata based on the original sample design, using the PSUs with responding sample. For states with two sampled PSUs in which data was collected in both, the two PSUs were paired to form a variance stratum. Otherwise, we grouped a PSU with another PSU (or two other PSUs) within the same region based on the similarity of the Cycle 1 small area estimate of the percentage at or below Level 1 in literacy. This resulted in 14 variance strata.

[3] See https://www.oecd.org/en/data/datasets/PIAAC-2nd-Cycle-Database.html

We then formed replicate weights for the supplemental sample by treating the core and supplemental samples as independent samples from the same population. Given that the core sample PSUs had had no chance of selection in the supplemental sample, independence does not hold, but this approach should provide a conservative estimate of variance. Based on the 14 variance strata for the supplemental sample, we created 44 replicates using Fay's method (k = 0.3) and a Hadamard matrix of size 44. In combining the samples, the total number of replicates remained at 44, as each of the 44 supplemental sample replicates was paired at random with one of the 44 core sample replicates. All weighting adjustments applied to the full sample weights were also applied to the replicate weights.

4. Evaluation and Outcomes

4.1 Evaluation of National Estimates

NCES and OECD required an evaluation to determine if the supplemental sample could be included in national and international reports. For the evaluation, we produced weights for the core-only sample and compared the results against the combined sample in terms of the weights, estimates, variances, and associations. The weighting process for the core-only sample was the same as that for the composited sample, except without the pre-compositing calibration and compositing steps.

For cases in the core sample, the correlation between the core-only weights and combined sample weights was 0.989. The strong relationship between the two can be seen in Figure 4.

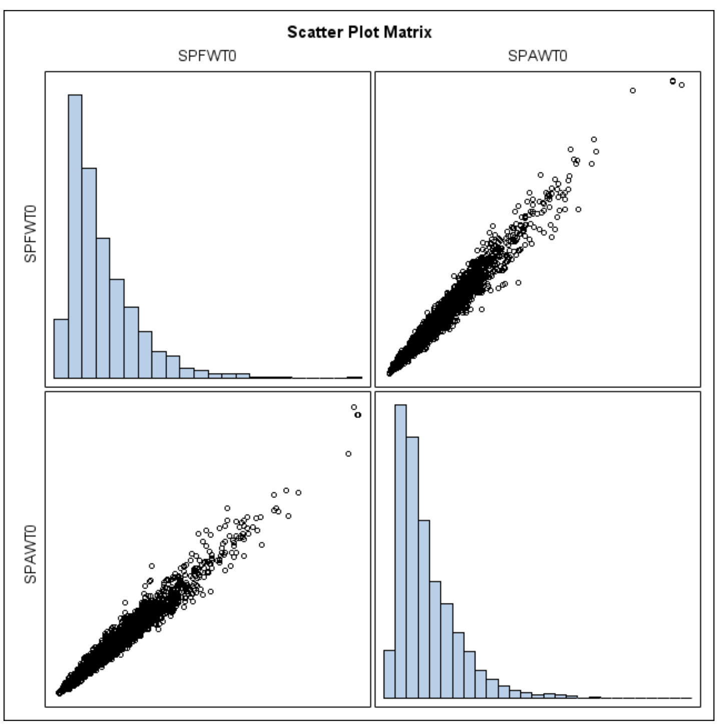


Figure 4: Scatterplot matrix comparing the core-only weights (SPAWT0) to the combined sample weights (SPFWT0) for the core sample cases

For evaluating estimates, we compared weighted proportions of survey variables. One set of survey variables were used in nonresponse adjustments – MSA status, literacy-related status, presence of children in the household, and age. The other set of variables for the analysis came from the Background Questionnaire, were not used in weighting adjustments, and were believed to be related to proficiency. This included employment status, computer experience, language spoken at home, and financial literacy. The proficiency scores were not available at the time of the evaluation. We performed t-tests on the difference between the core-only estimate and combined sample estimate, and none were significantly different at $\alpha = 0.05$. Figure 5 plots the core-only estimates versus the combined sample estimates, and it is evident that the two sets of estimates are very similar.

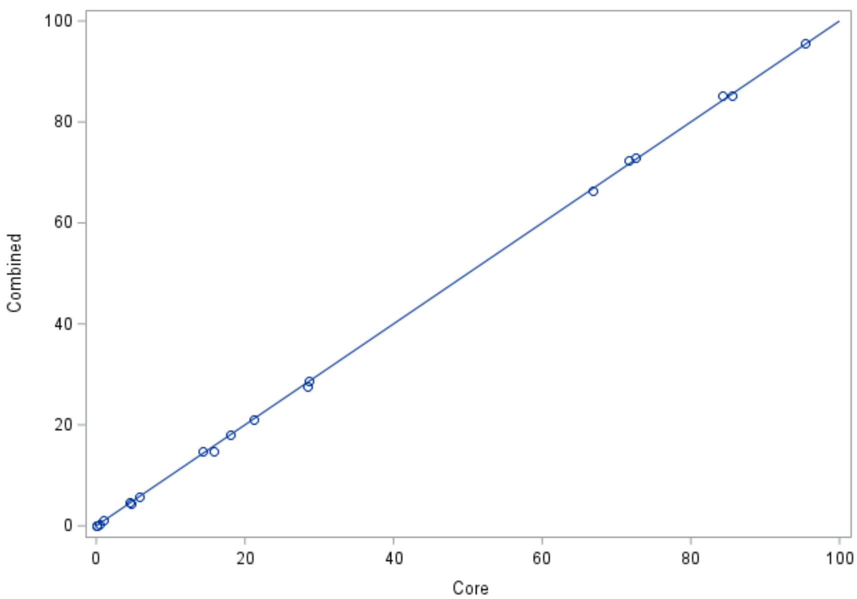


Figure 5. Background Questionnaire variable estimates (in percentages) for the core-only sample versus the combined sample

We also compared the standard errors of the same set of BQ estimates. No statistical test was performed, but from Figure 6, the standard errors can be seen to be of similar magnitude. As noted above, the method for combining the samples was partly chosen to limit the variation in the weights. This was accomplished, and the two sets of weights had similar CVs – 69% for core-only and 68% for combined.

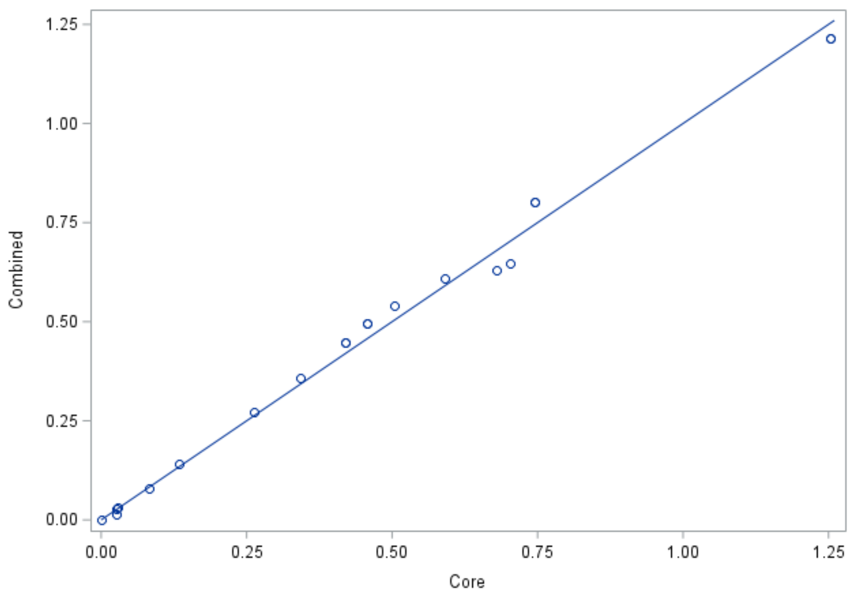


Figure 6. Standard errors of Background Questionnaire estimates for the core-only sample versus the combined sample

We used chi-square tests and regression to evaluate the effect of the combined sample on associations. First, we ran chi-square tests of associations between auxiliary variables, such as education and race/ethnicity. Among 21 tests, there was agreement on 20 as to whether the association was statistically significant at $\alpha = 0.05$. Next, we fit a regression model for education using three auxiliary variables (age, dwelling unit type, and the percentage of the population age 25 and older with a high school education) as predictors. The R-squared of the model was 11% when fit with either the core-only sample or the combined sample. The three auxiliary variables were significant in both models.

The results are not surprising given the size of the supplemental sample, but the evaluation provides further confirmation of the quality of the combined sample.

4.2 Modeling

A main advantage of incorporating the supplemental sample was that the combined sample met the Consortium's minimum sample size requirements for the IRT model. The sample was deemed sufficient to evaluate the quality of items and detect any country-by-language misfit at the national level. The resulting data was of high quality, meaning results from the United States could be included in international reports.

For small area estimation, the larger sample size should allow for less reliance on the model. Although incomplete, the supplemental sample helped with the goal of increasing the number of states and counties with data. The number of states with data increased from 34 for the core-only sample to 48 for the combined sample, and the number of counties with data increased from 89 to 126. In addition, the number of states with two or more PSUs increased from 31 to 37. As noted above, having two or more PSUs allows for direct estimation of variance for input into the models and captures some of the diversity of counties within a state. The small area estimation work is on-going, and no results are available at the time of this paper, but it is expected that the larger sample will help improve the quality of the estimates.

5. Conclusions

The challenge faced by U.S. PIAAC was one of an incomplete supplemental sample, i.e., a supplemental sample that began as a probability sample but where data collection was stopped short. While the selection probabilities were initially calculatable for the supplemental sample, they were unknown due to field staff working their assigned (non-random) subsets of the full supplemental sample in the first two months of the data collection period. Some alternative methods use a model-assisted approach with the survey outcomes and available auxiliary data; however, survey outcomes were not available at the time of weighting. Therefore, we chose a calibration and compositing approach to combine the incomplete sample and the probability reference sample, with the aim of reducing bias in the incomplete sample while limiting the variance. Unlike the typical situation, the non-probability sample was much smaller than the probability sample, and so inclusion bias in the non-probability sample was less of a concern. We strove to limit the bias as much as possible by calibrating the sample on variables related to inclusion and the survey outcome. We concluded that including the incomplete sample had a negligible effect on national estimates and should help strengthen model-based estimates by increasing the number of responses per assessment item for IRT models and the number of states and counties with data for small area estimation models.

References

Chakravarti, I.M., Laha, R.G., and Roy, J. (1967). *Handbook of Methods of Applied Statistics, Volume I*, John Wiley and Sons, pp. 392-394.

Dever, J. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*.

Elliott, M. and Valliant, R. (2017). Inference for nonprobability samples. *Statist. Sci*. 32 (2) 249 - 264, May 2017. https://doi.org/10.1214/16-STS598.

Judkins, D. R. (1990). Fay's Method for Variance Estimation. *Journal of Official Statistics*, 6, 3, 223-239.

Krenzke, T., and Haung, W. (2009). Revisiting Nested Stratification of Primary Sampling Units. *Proceedings of the Federal Committee on Statistical Methodology.* https://nces.ed.gov/FCSM/pdf/2009FCSM_Krenzke_IX-C.pdf.

Krenzke, T., VanDeKerckhove, W., Thornton, N., Diaz-Hoffmann, L., Hogan, J., Mohadjer, L., Li, L., Li, J., Yamamoto, K., Khorramdel, L., and Ali, U. (2019). *U.S. Program for the International Assessment of Adult Competencies (PIAAC) 2012/2014/2017: Main Study, National Supplement, and PIAAC 2017 Technical Report* (NCES 2020-224). U.S. Department of Education. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020224.

Krenzke, T., and Mohadjer, L. (2020). Application of probability-based link-tracing and non-probability approaches to sampling out-of-school youth in developing countries. *Journal of Survey Statistics and Methodology*. doi: https://doi.org/10.1093/jssam/smaa010

National Center for Education Statistics. (2012). *2012 NCES Statistical Standards*. Available at https://nces.ed.gov/statprog/2012/.

Organisation for Economic Cooperation and Development. (2022). *Cycle 2 PIAAC Technical Standards and Guidelines*. https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/piaac/technical-standards-and-guidelines/cycle-2/PIAAC_CY2_Technical_Standards_and_Guidelines.pdf/_jcr_content/renditions/original./PIAAC_CY2_Technical_Standards_and_Guidelines.pdf.

PIAAC International Highlights Web Report (NCES 2020-127). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Available at https://nces.ed.gov/surveys/piaac/international_context.asp.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263. DOI: 10.1093/jssam/smz003.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm.

Yamamoto, K., Khorramdel, L., von Davier, M. Ali, U., and Robin, R. (2019). Scaling PIAAC cognitive data. *Technical Report of the Survey of Adult Skills (PIAAC) (3rd Edition)*, Chapter 17. Organisation for Economic Cooperation and Development. https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/piaac/technical-reports/cycle-1/PIAAC_Technical_Report_2019.pdf.