Westat®

Improving lives through research™

Programme for the International Assessment of Adult Competencies

Small Area Estimation Research

Tom Krenzke, Leyla Mohadjer, Jane Li, Wendy Van de Kerckhove, Lin Li, Weijia Ren, and Henok Adbaru

September 2018

Prepared for:
Organisation of Economic Co-operation and Development
2, rue André Pascal
75775 Paris Cedex 16
France

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

This page is intentionally blank

Table of Contents

1. Introduction

The Programme for the International Assessment of Adult Competencies (PIAAC) sample is designed to produce internationally comparable and nationally representative direct estimates (based solely on survey data) with adequate levels of precision for the nations as a whole and for major population subgroups. However, the Organisation for Economic Cooperation and Development (OECD), and several of the participating countries in Cycle 1 of PIAAC, have expressed interest in using PIAAC data to create proficiency estimates for local areas where PIAAC sample size is too small (or equal to zero) to produce any direct estimates. Small area estimation (SAE) methods facilitate the estimation of the proficiency distribution in subpopulations not initially targeted in large scale surveys.

A considerable amount of research and development in small area estimation (SAE) methods has taken place since the text by Rao (2003), which presents a comprehensive overview of the methods, history, and applications of SAE[1] methods. The book has since been updated (Rao and Molina, 2015), and much research and development activity has been on-going on this topic in recent years. The development of SAE approaches has made it possible to meet the growing demands for more information at lower levels of geography. It is no different for PIAAC. The application of SAE approaches to PIAAC data may provide an affordable option for countries to produce indirect estimates for their small areas of interest.

This paper summarizes the research results from applying SAE methods using PIAAC data from five countries that participated in Cycle 1, with various core national sample designs. First, the paper provides some background on PIAAC (Section 1.1), SAE techniques (Section 1.2), goals of this research (Section 1.3), the selection of countries for this research (Section 1.4), and the main steps included in this research (Section 1.5). Section 2 contains a description of the process of involving a selected group of countries, as well as the standard guidelines developed for requesting data from the countries. In addition, this section includes a brief description of issues that arose during the data submission process. The remainder of the section is devoted to brief descriptions of the sample

[1] Note that in this document, "SAE" refers to small area estimation, and small area estimates, interchangeably.

designs, sample sizes and small areas, and the auxiliary variables (covariates) for each country. In Section 3, the direct estimation process is discussed, which includes the use of the survey regression estimator and variance smoothing.

Section 4 introduces the models and the concept of accounting for various sources of error that contribute to the precision of the resulting estimates. The work includes an evaluation of methodology and approaches, including both unit-level and area-level SAE modeling. Section 5 provides results of the evaluation of the SAEs and precision measures, and Section 6 summarizes the overall outcomes, including feedback from countries, and a conclusion on critical factors to be considered. Finally, Section 7 provides some general thoughts toward producing publishable SAEs for PIAAC countries in general, including issues around sharing data at small area levels, and the suitability of the PIAAC country national samples for small area estimation.

We refer to this work as Phase 1 research -- using a group of countries to explore what models are suitable for different countries. Phase 2 (production phase) would focus more on each country individually, and involve more work in selecting covariates and performing model diagnostics to help evaluate various sets of SAEs toward publishable estimates. The Phase 2 analysis is expected to use the methodology most appropriate for each country's data and sample design.

The authors wish to acknowledge the use of some information that has been gained and gathered under contract to the U.S. National Center for Education Statistics. In addition, some of the general SAE text in this report was also originally developed for the analysis plan for the U.S. PIAAC. The authors are grateful to the countries for the help and participation in this research. In addition, the authors acknowledge the invaluable insights and guidance provided by Bob Fay and Jon Rao throughout the small area research. Their vast experience in small area estimation was instrumental in resolving the various challenges faced during the research and development process. Lastly, the authors express appreciation for the opportunities to study this topic and apply the methods to better inform local areas about their proficiencies in literacy.

1.1 Programme for the International Assessment of Adult Competencies

PIAAC is a multicycle survey of adult skills and competencies sponsored by the OECD. The survey examines a range of basic skills in the information age and assesses these adult skills consistently across participating countries. The first cycle of PIAAC includes three rounds: 24 countries participated in 2011–12 (round 1); 9 additional countries participated in 2014–15 (round 2); and 5 additional countries are participating in 2017–18 (round 3). In general, the sampling goal was to achieve 5,000 completed assessments, which included the following three domains:

- Literacy (including reading component);
- Numeracy; and
- Problem-solving in technology rich environment.

The sample designs varied across countries. Because of the need to conduct the assessment in-person, most countries chose to cluster their sample into Primary Sampling Units (PSUs) to reduce costs of interviewing within households. Inherent in PIAAC is both informative sampling (clustering and differential base weights) and informative nonresponse (non-ignorable proficiency-related nonresponse), as evident by the steps included in the weighting process which accounts for differential probabilities of selection, nonresponse adjustments, and calibration of the weights. Both informative sample design and nonresponse should be taken into account when generating SAEs.

The test design for PIAAC is based on a variant matrix sampling (OECD, 2016) where each respondent was administered a subset of items from the total item pool. Therefore, item response theory (IRT) scaling was used to derive scores for each domain. To increase the accuracy of the cognitive measurement, PIAAC uses plausible values (multiple imputations) drawn from a posterior distribution by combining the IRT scaling of the cognitive items with a latent regression model using information from the background questionnaire (BQ) in a population model.

1.2 Small Area Estimation

The essence of SAE is to use covariates at the small-area level in combination with survey data to model the small area parameters of interest. As the demand for reliable small area estimates has greatly increased in the past decades, the SAE literature and research findings also has grown rapidly.

This was evident for example when the International Statistical Institute (ISI) held its 61st ISI World Statistics Congress Satellite Meeting on small area estimation in July 2017 in Paris, France.[2] The main purpose was to assess the current state of development and usage of small area methodology and to discuss upcoming challenges and solutions. In recent years, the ideas of SAE have been associated with and applied to other fields such as Big Data, confidentiality protection, and record linkage. This meeting served as a bridge between statisticians and practitioners working on SAE in academia, private and government agencies, and other fields.

Section 4 describes the various approaches developed under SAE methodology and used in this research. In general, there are two major types of models: area level and unit level models. The area-level approach models the small area parameter of interest in terms of covariates at the area-level, whereas the unit-level[3] approach models the underlying variable of interest in terms of unit-level covariates known at the small-area level, and then aggregating the individual predictions for each small area.

Specific traditional SAE methodologies include Hierarchical Bayes (HB), which is applied when there is a fixed prior distribution, and Empirical Bayes (EB) is conducted when the prior distribution is based on the data itself. Empirical Best Linear Unbiased Predictor (EBLUP) can be used to estimate random effects, such as through SAS Proc Mixed. The Linear Mixed Model (LMM) is when the dependent variable follows a normal distribution. The Generalized Linear Mixed Model (GLMM) is considered when the dependent variable has a nonnormal error distribution. The reader can find many details on the various types of SAE models in Rao and Molina (2015).

Pfefferman (2013) reviews and discusses some of the important new developments in small area estimation method since the publication of Rao's *Small Area Estimation* in 2003. The review covers both design-based (frequentist) and model-dependent (Bayesian) methods. Pfefferman also reviews the new developments on SAE under informative sampling and nonresponse as well as model selection and checking. In the case of informative sampling or not-missing-at-random nonresponse, the model assumed for the population may not apply to the sample data. If not properly accounted

[2] Presentations slides available for some presentations at http://sae2017.ensai.fr/presentation.

[3] "Unit level" can mean at the individual sample unit level (person or household), or it could mean a geographic area lower than the small area.

for, the predictions can be seriously biased. He concludes that model-based predictors are generally more accurate and permit predictions for nonsampled areas, where the design-based theory does not exist. Compared to the frequentist approach, the Bayesian approach is more flexible and easier to handle inference.

The following paragraphs provide a brief review of two SAE papers/research activities related to PIAAC.

Bijlsma, van den Brakel, van der Velden, and Allen (2017) used the Netherlands' PIAAC data and focused on obtaining the literacy estimates at the municipality level in the Netherlands using model-based SAE techniques in an HB framework. The Netherlands participated in round 1 of PIAAC and achieved a total sample size of about 5,000, and less than 20 observations in most municipalities. Direct estimates using the observations from the domains can have unacceptably large design variances. To increase the precision of municipal estimates, small area models are applied to increase the effective sample size of each municipality. Two literacy measures were of interest per area: the average literacy score and the proportion of low literates. A basic unit-level model originally proposed by Battese, Harter, and Fuller (1988) was used to model the average literacy score since literacy scores are continuous per individual and areas are assumed to have a linear relation with individual-level covariates. In the case of the proportion of low literates, the dependent variable is dichotomous at the individual level, equal to one if the score is below the low-literacy cutoff point of 226 and zero otherwise. Therefore, an area-level model originally proposed by Fay and Herriot (1979) was used to model the proportion of low literates. The models assume that the effects of covariates at the area level are the same as at the national level, with random effects capturing regional differences.

Yamamoto (2014) presented an approach to produce the estimates of skill distribution for provinces based on population parameters derived from the Canadian Program for the International Assessment of Adult Competencies (PIAAC) data and covariate information such as census. The models assumed the similarities of the covariance structure among skill and background variables between the population and subpopulation. As a result, the distributional characteristics of a skill variable could be derived from the respondents. The population model used for PIAAC is a combination of an item response theory (IRT) model and a latent regression model. Once the population parameters are estimated from the respondents, plausible values are drawn from

posterior distribution. The paper evaluates the impact of reduced background information and reduced responses on cognitive items on the accuracy of the estimation. The results of this type of unit-level approach show that the population parameters based on the national data captures significant information regarding the skill distribution even with a relatively small number of background variables.

There must be careful attention to further understand the impact of informative sampling and informative nonresponse, especially for unit-level models. For example, if clusters were selected, or exist, within small areas, the unit-level models do not address the impact of such clustering on the variance estimates.

The approach may also need modification to account for measurement error if the covariates information is unstable. For example, because a full cross-tabulation of several predictors is needed for a unit-level non-linear model, the estimates for each cell of a full crosstabulation of six variables will be unstable. In order to allow for the measurement error to propagate through to the small area estimates, an approach such as introduced by Ybarra and Lohr (2008) could be incorporated; however it may not be able to extend to many variables.

1.3 Goals of Research

The main purpose for this research is to evaluate various SAE approaches across countries of different sizes and with different PIAAC sample designs toward developing an understanding, and guidance, on how SAE can be implemented for PIAAC.

Types of Models

Both major types of models have been developed for SAE: 1) area-level, and 2) unit-level. Depending upon the fit, the final country models could be linear or non-linear. As discussed in Rao and Molina (2015), the area-level approach includes a sampling model and a linking model, where the two models can be matched for linear estimation, and unmatched for nonlinear estimation. The unit-level model takes advantage of powerful covariates at the person level, if the data exists and can match in definition to the PIAAC data. Both types of models, as applicable, were fit to data from

each country. Further detail on model descriptions and applications to country data are provided in Section 4.

Types of Estimates

In this effort, we were interested in producing SAEs of adults at the lower literacy levels, specifically, the proportion in Level 1 or below. In addition, we included statistics on average literacy scores (mean values) in our research to fully examine and evaluate various methods and models. Another benefit of looking at averages along with proportions at the lower literacy levels is the ability to provide a better picture of literacy in local areas, as described by Bijlsma, et al. (2017). For example, different mean score literacy estimates in two small areas with the same proportion at or below Level 1 points to the differences between the distribution of the adult population at the higher literacy levels in the two small areas.

1.4 Participating Countries

In collaboration with OECD, a handful of countries with diverse sample designs and various levels of access to covariates (as observed through the weighting and nonresponse bias activities) were selected to participate in this research. The countries were recruited by the OECD. Table 1-1 provides the number of sampling stages, type of frame and population size for each country.

Table 1-1 Number of sampling stages, frame type, and population size for participating countries

Country	Number of sampling stages	Frame type	Population size
Germany	2	Registry	82.5 million
Italy	3	Screener	60.6 million
New Zealand	4	Screener	4.7 million
Slovakia	2	Registry	5.4 million
Sweden	1	Registry	9.9 million

Source of population size: 2016 World Bank. https://data.worldbank.org/indicator/SP.POP.TOTL

The authors acknowledge our primary contacts for each country and appreciate the timely responses to our questions and the feedback they provided.

- Germany: Silke Martin, Stefan Zins, Beatrice Rammstedt

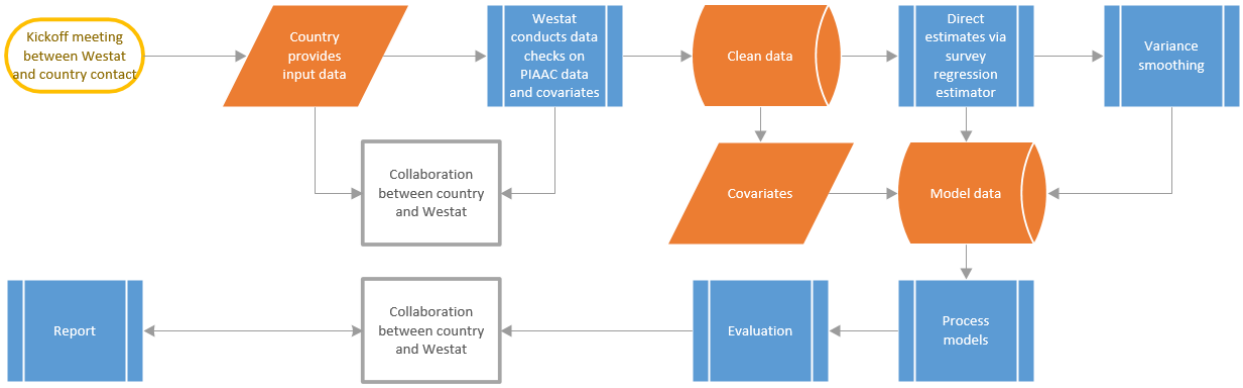
- Italy: Simona Mineo, Valentina Gualtieri
- New Zealand: David Earle, Paul Satherley
- Slovakia: Ildiko Pathoova
- Sweden: Lotta Larsson, Johan Lofgren

1.5 Main Steps

This initial phase of research included the following main steps:

- Kickoff meetings occurred with each country. During the meeting, we provided an overview of the research goals, SAE methods, estimates of interest, geographic areas of interest, key covariates, and the flow of collaboration with the countries (shown in Figure 1-1). Much of the discussion centered on the input data request (discussed in Section 2). The meetings concluded with a brief summary of the potential SAE models that were being developed.
- Data cleaning. A standard format was provided to each country for the transfer of the input data for the SAE processing. Once received, there were several checks that were conducted to ensure a clear understanding of the categories of each variable received and that the data were as expected. This was a time consuming process. More discussion on this topic is given in Section 2.
- Direct estimates. Simple weighted estimates and variance estimates were produced for each country, with exception to Germany who provided the estimates. In addition, survey regression estimates were produced to align with covariate estimates for each small area. Variance estimates were then smoothed to help satisfy a key assumption in SAE modeling, which is to assume the variances are known.
- Generate model results. A different set of SAE models were chosen for each country depending on the nature of their sample and their covariate data. The SAE models were processed for the literacy component proportion at or below Level 1, and for the average.
- Model evaluation. Limited model diagnostics were performed. Several plots were generated to review the performance of point estimates and mean square error.
- Country feedback. The model results were shared with countries for their feedback.

Figure 1-1 Collaboration between the country and Westat during the SAE initial research process



2. Country Data

After the initial consultations, the first step in the SAE research process was to obtain the required data from each country. Section 2.1 describes the challenges in accomplishing and the effort put into obtaining and processing the initial data files. The SAE procedure can vary depending on the country's sample design, the definitions and sample sizes of the small areas, and the available covariates. Details on these topics are provided in Sections 2.2, 2.3, and 2.4, respectively.

2.1 Data Request

To process the small area estimation models, each country was asked to provide two files according to the layout given in Appendix A: a PIAAC data file and a population file. The PIAAC data file was to include the following variables for each PIAAC respondent: person identifier, small area (SA) identifier, variance cluster identifier, final full sample and replicate weights, literacy scores (10 plausible values), and covariates. The population file was to include the covariates for the universe of persons for each SA. If the country was not able to include population totals for a full crosstab of the covariates by SA, arrangements were made for countries to provide frequencies or partial cross-tabulations (e.g., involving 1 to 3 variables). We should note that for fitting a unit-level model, the covariates on the PIAAC data file should have the same coverage, definitions, and categories as those on the population file.

Westat reviewed and processed the files for input to the SAE process, as described in Section 2.1.1. Countries were generally able to provide the requested information, but there were some limitations to the data, as explained in Section 2.1.2. Section 2.1.3 summarizes lessons learned in acquiring data files for SAE from various countries.

2.1.1 Data Processing

Using the input data from each country, Westat created three files:

- The respondent-level file serves as input to direct estimation, smoothing, and unit-level modeling. It was based on the country's PIAAC data file, with dummy variables derived for each covariate.

- The crosstab file is used in unit-level modeling. It includes the crosstab of covariates in the country's population file(s) for each SA.
- The SA-level file is used in area-level models and for the SRE[4]. It was created from the country's population file, and it includes means or proportions of each covariate for each SA.

Given the constraints in the submitted data, it was not possible to produce a respondent-level file for Germany or crosstab files for Germany and Sweden. More information is provided in Section 2.1.2. Table 2-1 shows the files produced for each country.

Table 2-1 Input files to the SAE process

Country	Respondent-level file	Crosstab file	SA-level file
Germany			✓
Italy	✓	✓	✓
New Zealand	✓	✓	✓
Slovakia	✓	✓	✓
Sweden	✓		✓

Westat created standardized variable names across files and across countries. In some cases, covariates with a large number of categories were collapsed into fewer categories, and this was done consistently across files. We also reviewed the definitions and mean values of the covariates on the PIAAC data file and the population file to determine whether the variables were consistent between the two files. For the PIAAC data file, means were calculated using final weights.

2.1.2 Data File Constraints and Limitations

Generally, countries were able to provide the requested information, or it could be obtained from internal Consortium files from the weighting process, sample design international file (SDIF), or public use file (PUF). However, there were limitations to the data based on confidentiality, file layout compliance, and matching covariate definitions.

[4] The SRE is used to smooth direct estimates for use in area-level models.

Confidentiality

Germany and Sweden faced confidentiality restrictions in providing microdata. Germany could not provide the SA identifier for each respondent and thus supplied PIAAC data summarized to the area-level. Without the respondent-level data, it is not possible to fit a non-linear unit-level model or an SRE. Sweden had some interest in producing small area estimates for the 21 counties in Sweden but could not provide the microdata at this level. They opted to use the eight NUTS2 areas identified on the PIAAC public use file. Given the small number of SAs, a model-assisted direct estimation approach was conducted, as well as a unit-level EBLUP model.

Additionally, New Zealand's population file included an extra level of processing because some counts were suppressed for confidentiality reasons. Westat imputed counts for these cells before creating the crosstab and SA-level files.

File Layout Compliance

Ideally, we would like to receive a full crosstab of covariates on the population file to facilitate unit-level modeling with non-linear relationships. However, it is not a requirement for the small area estimation process. Italy and New Zealand were the only countries with a full crosstab of all covariates. In New Zealand's crosstab file, persons could be counted in more than one ethnicity category, and thus the sum over the categories was greater than the population. This complexity was overlooked in processing the files, and persons with multiple ethnicities are double-counted in the numerator and denominator of SA-level covariates. This should have minimal impact on the Phase 1 results. Slovakia provided partial crosstabs. Germany provided an area-level file that had area-level estimates for covariates, and the SAs could not be linked to their PIAAC data because of the confidentiality restrictions mentioned above, and therefore no unit-level modeling could be done. Sweden's SAs were readily available with the PIAAC microdata from the PIAAC Round 1 work and therefore unit-level linear modeling with area random effects could be done.

Matching Definitions

Covariates that are available on the PIAAC data file and population file can be used in unit-level models, but only if covariates are defined consistently between the two files. The full list of

covariates supplied by each country is given in Section 2.4. The usefulness of the covariates was limited for the following reasons:

- The covariate was available in the population file but not the PIAAC data file.
- The definitions differed, e.g., the categories of "current economic activity" for Slovakia did not exactly align with the employment status categories in PIAAC.
- The population totals were not provided for the exact PIAAC target population. Specifically, New Zealand and Slovakia's population totals were for the population age 15 to 64 rather than 16 to 65.
- The definitions and target population appeared to be consistent (i.e., the first two bullets were met), but the mean value of the covariate differed between the two files by more than two percentage points.
- All conditions described in the above bullets were met, but the SRE produced unusual results, specifically the SREs were consistently higher than the direct estimates or consistently lower than the direct estimates, and removing the covariate addressed this issue. This indicates that there is likely an unidentified inconsistency in the covariate definition between the two files.

Variables determined to be inconsistent between the two files were excluded from unit-level models and only considered in area-level models. The slight age range discrepancies for New Zealand and Slovakia were considered to have minimal impact for the Phase 1 research, and thus we did not exclude covariates on this basis.

2.1.3 Lessons Learned

Some improvements could be made to the data request and processing steps to reduce processing time and increase the likelihood of receiving useful data. For Phase 2 implementation, we recommend countries to conduct the following steps before submitting their data to Westat:

- Create a mapping of the PIAAC covariates to the population covariates.
- Collapse covariates so that they have a maximum of three or four categories.
- Use consistent variable names and file layouts.
- Review definitions and distributions of covariates in the PIAAC data file and population file for consistency.

2.2 Sample Designs

Table 2-2 summarizes the sample designs and sample sizes for the five participating countries, followed by further details provided below. All countries but Sweden had clustered samples, with between 260 and 1,000 units at the first stage. The final sample sizes ranged from 4,469 to 6,177.

Table 2-2 Sample designs

Country	Sample design	Number of sampled PSUs	Number of completes
Germany	2-stage cluster sample	277	5,465
Italy	3-stage cluster sample	260	4,621
New Zealand	4-stage cluster sample	1,000	6,177
Slovakia	2-stage cluster sample	562	5,723
Sweden	1-stage sample	Not applicable	4,469

Germany. The sample design for the 2012 PIAAC survey was comprised of a two-stage cluster sample. The first stage included a sample of 277 communities from many strata using region and urban/rural status. Controlled rounding methodology was used in the selection of the communities. The second stage included two-phases. Phase 1 involved asking communities to select and EPSEM sample of individuals from their local registry. Then in Phase 2, within each community, the individuals selected in Phase 1 were allocated to a matrix that was divided into six age groups and gender categories. 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.

Italy. The sample design for the 2012 PIAAC survey was comprised of a three-stage cluster sample. The first stage included a sample of 260 area primary sampling units (PSUs) selected with probabilities proportionate to size, sorted by total population within explicit strata based on equal sized regions. In the second stage, a frame of dwelling units was formed from the registry. A total of 11,592 dwelling units were selected within sampled PSUs. One person from each DU was pre-selected from the DU registry. A screener questionnaire was administered to selected DUs. If the household composition was found to be different from the registry, persons were sorted by gender and age and the selection grid was used.

New Zealand. The sample design for the 2014 PIAAC survey was comprised of a four-stage cluster sample. The first stage included a sample of 1,000 area clusters (PSUs) selected with probabilities

proportionate to the number of occupied dwelling units (and units under construction) sorted by total population within explicit strata based on equal sized regions. In the second stage, 1,000 meshblocks were selected using the same size measure as PSUs, one from each PSU. In the third stage, the frame of dwelling units was sorted by geography and 16,392 dwelling units were selected within sampled meshblocks. A screener questionnaire was administered to selected DUs and one person was selected per DU.

Slovakia. The sample design for the 2012 PIAAC survey was comprised of a two-stage cluster sample. The first stage included a sample of 562 municipalities (PSUs) selected with probabilities proportionate to the number of adults 16 to 65 years old, sorted by total population within explicit strata based on region and municipality size. In the second stage, within PSUs, persons on the population registry were sorted by gender and age and selected using a systematic random sample.

Sweden. The sample design for the 2012 PIAAC survey was comprised of a one-stage simple random sample within explicit strata. Strata were formed from gender, age, country of birth and level of education.

2.3 Defining Small Areas

For SAs, following OECD's recommendations, Westat suggested NUTS level 2 for larger countries and NUTS level 3 for smaller ones, although the final choice was made by countries depending on their interest. The SA definitions for each country are given in Table 2-3. In all five countries, the SAs are larger areas than the PSUs, meaning that the sample is clustered within an SA. The number of SAs varies from eight for Sweden to 110 for Italy. Germany, Slovakia, and Sweden have PIAAC sample in all areas. Italy and New Zealand have sample in over 80% of areas. In addition, the sample size within an SA varies. For Germany, Italy, and New Zealand, the majority of SAs have between 31 and 100 completed cases. For Slovakia, over 50% have over 100 completed cases, and for Sweden, all SAs have a sample size over 100.

Table 2-3 Small area definitions, population counts, and sample sizes

Country	Small area (SA) description	Number of SAs	Number of SAs with sample	Number of SAs with n =		
				1-30	31-100	101+
Germany	Collapsed spatial planning regions	85	85	12	60	13
Italy	Provinces	110	91	35	50	6
New Zealand	Territorial Authorities/ Community Boards	87	84	21	47	16
Slovakia	Districts/counties (LAU_1)	79	79	15	19	45
Sweden	NUTS2	8	8	0	0	8

2.4 Covariates

The covariates in the SAE models should be highly predictive of the SA estimates of interest. Based on this criterion, Westat recommended obtaining population data on age, gender, race/ethnicity, education attainment, employment status, poverty status, and foreign-born status. In addition, the population totals should come from a population census, administrative data, or a large national survey. For fitting a unit-level model, the covariates on the PIAAC data file should have the same coverage, definitions, and categories as those on the population file.

Table 2-4 shows the covariates chosen by each country. All countries included variables related to gender, age, nationality, and education, and all but Sweden included an employment-related variable. Additional covariates were household size, marital status, ethnic group, and language. The population totals generally came from a census or registry from the same period as the PIAAC survey. The exceptions are Germany's population totals, which are estimates from a survey (the Germany Micro Census). Westat found the covariates to have a weak association with the direct estimates, and Germany is looking into whether the variances of the covariate estimates are large.

The table also shows the number of levels for each covariate, as provided by the country. In general, Westat collapsed variables to four or fewer levels for use in the SAE process. Some covariates were only available on the population file and did not have an equivalent variable in the PIAAC data. In addition, education and employment status were often found to match poorly between the two files, based on the criteria in Section 2.1.2. Such covariates can be used in area-level models only. The covariates with consistent definitions between the PIAAC and population files are indicated in bold.

Table 2-4 Covariates on country input files

Country	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Source of population totals
Germany	Gender*	Age* (4 levels)	Nationality* (3 levels)	Educational attainment* (5 levels)	Employment Status* (3 levels)			Micro Census (2011)
Italy	**Gender**	**Age** (Exact age)	Citizenship* (2 levels)	**Educational attainment** (6 levels)	**Employment Status** (7 levels)	**Number of people in household** (5 levels)	Marital status* (6 levels)	Census (2011)
New Zealand	**Gender**	**Age** (6 levels)	**Birthplace** (2 levels)	Highest qualification (4 levels)	Work and Labor force status (2 levels)	Ethnic Group (3 levels)		Census of Population and Dwellings (2013)
Slovakia	**Gender**	**Age** (21 levels)	Nationality* (16 levels)	Highest education (9 levels)	Economic activity (13 levels)	Language spoken at home* (14 levels)		Population Census (2011)
Sweden	**Gender**	**Age** (5 levels)	**Birthplace** (2 levels)	Highest Education (4 levels)				Swedish register (2012)

* On population file only; not available on PIAAC data file.

NOTE: Bold font indicates consistent definitions between the PIAAC and population files, as defined in Section 2.1.2.

3. Direct Estimation

A large number of tasks were administered in the Program for the International Assessment of Adult Competencies (PIAAC) assessment to ensure the survey covered a broad range of proficiency tasks. However, to keep the testing time at a reasonable level, each participant was given a subset of the pool of literacy tasks using a matrix sample design in a way that ensured that each of the tasks was administered to a nationally representative sample of adults, with some core tasks being administered to all sampled adults. Because different respondents took different sets of items that could be of various levels of difficulty, it would be inappropriate to base the proficiency estimates simply on the number of correct answers obtained. Therefore, large-scale assessments using matrix sampling rely on item response theory (IRT) models. The PIAAC IRT modeling resulted in 10 plausible values (PV) for each respondent, reflecting the uncertainty in the respondents' proficiency estimate. More information can be found in OECD (2016).

Section 3.1 describes the steps taken to compute direct estimates of average literacy and the proportion at or below Level 1 in literacy, using the 10 literacy PVs. The variances of the direct point estimates and variance estimates can be large in SA's with small sample sizes. Section 3.2 describes efforts to improve the stability of the estimates using a survey regression estimator (SRE) and variance smoothing process. The resulting estimates serve as input to the SAE modeling.

3.1 Direct Estimates

To process the small area estimation models, we first produced direct point estimates and variance estimates for the proportion at or below Level 1 in literacy and average literacy. Variance estimates should account for the error associated with the IRT modeling in addition to the sampling error. To handle the plausible values properly, a multiple imputation (MI) approach, as shown in Rubin (1987) was used for calculating direct estimates and the associated variances.

Using the respondent-level file described in Section 2, we created direct estimates of the proportion at or below Level 1 in literacy for all SAs with at least one respondent at or below Level 1 for at least one literacy PV. This criterion was necessary for variances to be estimable. All SAs meeting this criterion had at least five respondents, so we did not make any further exclusions based on sample

size. We created direct estimates of average literacy for all SAs with at least five respondents. Respondents were defined as cases with a final weight, i.e., respondents to the background questionnaire (BQ) and sampled persons that did not respond to the BQ for a literacy-related reason (language barrier, reading/writing barrier, or mental disability). The latter group did not have literacy scores, but they were assumed to be at or below Level 1 for proportions, and for averages, we imputed a literacy score using the first percentile of respondent's scores. The resulting number of SAs receiving direct estimates is shown in Table 3-1.

Table 3-1 Counts of small areas by country

Country	SAs	SAs with sample	SAs with direct estimates for proportions	SAs with direct estimates for averages
Germany	85	85	85	85
Italy	110	91	90	91
New Zealand	87	84	81	83
Slovakia	79	79	77	79
Sweden	8	8	8	8

Note: The number of SAs used in the area models is the same as shown of averages and proportions, respectively.

To obtain the direct point estimates for Italy, New Zealand, Slovakia, and Sweden, the SA-level direct survey estimate was calculated for each literacy plausible value (*l*) using the Hajek estimator:

$$\hat{y}_{jl} = \sum_{k=1}^{n_j} w_{jk} y_{jkl} / \sum_{k=1}^{n_j} w_{jk},$$

Where

n_j = the number of cases in SA *j*,

w_{jk} = the final weight for person *k* in SA *j*,

$y_{jkl} = PVLIT_{jkl}$ for averages or I_{jkl} for proportions, where

$PVLIT_{jkl}$ = literacy plausible value for person *k* in SA *j*, where *l* = 1 to 10, and

$$I_{jkl} = \begin{cases} 1 \; if \; PVLIT_{jkl} < 226 \\ 0 \; otherwise \end{cases}.$$

A score of less than 226 is considered at or below Level 1. For literacy-related nonrespondents, the indicator I_{jkl} was set to 1 for all 10 plausible values.

We then applied the MI formulae to obtain the SA-level point estimate:

$$\hat{y}_j = \frac{1}{10}\sum_{l=1}^{10} \hat{y}_{jl},$$

And the variance:

$$\hat{\sigma}_j^2 = \hat{\sigma}_{Wj}^2 + (\frac{11}{10})\hat{\sigma}_{Bj}^2,$$

Where $\hat{\sigma}_{Wj}^2$ is the within-imputation variance and $\hat{\sigma}_{Bj}^2$ is the between-imputation variance. Variances were calculated using the final replicate weights and appropriate replication method for the country. We opted to use the replication method because we did not always have the necessary clustering information to implement Taylor Series. The variances for New Zealand employ delete-one jackknife (JK1) with 80 replicates. The variances for Italy, Slovakia, and Sweden use paired jackknife (JK2) with 80 replicates.

As described in Section 2, Germany was unable to provide respondent-level data due to confidentiality restrictions. Therefore, Germany computed direct point estimates and variances and included them on their SA-level delivery file. The file included two sets of direct variance estimates – one using Taylor series and one using the JK1 replication method with 80 replicates. For consistency with the other countries, Westat initially opted to use the JK1 estimates in the SAE process. However, this produced unexpected results, and so we switched to Taylor series based on Germany's recommendation.

3.2 Improved Direct Estimates

3.2.1 Survey Regression Estimator (SRE)

PIAAC was designed to be a nationally representative sample and does not produce efficient estimates at the SA-level. The survey regression estimator (SRE) is a model-assisted approach that is used to bring SA population estimates in line with external SA totals and improve the stability of the

survey estimates. The SRE process also helps to reduce variances that are used in the SAE modeling process. Rao and Molina (2015, pp. 21-23) describe the use of these estimates in small area estimation, their derivation, and the usual approach to estimating their variance. In addition to the x covariates available for the respondents, the values of the population totals $\mathbf{X}_j$ in SA j must be available for this estimator. For Germany, we did not have the covariates for the respondents, and thus no SRE was produced.

For the other countries, the SRE was derived for each plausible value as follows:

$$\hat{y}_{jl}^{surv} = \hat{y}_{jl} + (\bar{X}_j - \bar{x}_j)'\beta$$

where

$\hat{y}_{jl}$ = the survey estimate based on the *l*-th plausible value for SA *j*;

$\bar{X}_j$ = the vector of population means of the covariates;

$\bar{x}_j$ = the vector of sample means of the covariates; and

β = the vector of regression coefficients from the regression model of the relationship between y and *x*.

The covariates were limited to variables that were defined consistently for the respondents and the population, as described in Section 2. The list of covariates used in the SRE model for each country is given in Table 3-2. The table also indicates the strength of the covariates, as measured by the pairwise correlation with the direct estimates. Italy has a larger number of covariates, with correlations as high as 0.42. On the other end, Slovakia's SRE was limited to using age and gender, with correlation as high as 0.46.

Table 3-2 List of covariates for the SRE and strength of covariates

Country	Covariates (correlation[1] with direct estimate in parentheses)					
Italy	Gender, 1 level (0.02)	Age, mean (-0.32)		Education attainment, 2 levels (0.41, -0.31)	Employment status, 1 level (-0.36)	Number of people in the household, 2 levels (-0.42, 0.29)
New Zealand	Gender, 1 level (0.05)	Age, 4 levels (0.27, 0.06, -0.26, -0.17)	Birthplace, 1 level (0.02)			
Slovakia	Gender, 1 level (-0.32)	Age, 4 levels (0.46, -0.08, -0.14, -0.15)				
Sweden	Gender, 1 level, (0.43)	Age, 4 levels (-0.14, 0.23, 0.12, -0.28)	Birthplace, 1 level (-0.24)			

[1] As pointed out in Lahiri and Suntornchost (2015) the true population correlations are higher. The correlation estimates could be improved if the sampling error is taken into account, as described in Lahiri and Suntornchost (2015).

NOTE: No SRE was produced for Germany because Germany could not provide respondent-level data.

We then applied the MI formulae to produce the overall SRE estimate as:

$$\hat{y}_j^{surv} = \frac{1}{10}\sum_{l=1}^{10} \hat{y}_{jl}^{surv},$$

and the variance as:

$$\hat{\sigma}^2_{j(SRE)} = \hat{\sigma}^2_{Wj(SRE)} + (\frac{11}{10})\hat{\sigma}^2_{Bj(SRE)},$$

Where $\hat{\sigma}^2_{Wj(SRE)}$ is the within-imputation variance and $\hat{\sigma}^2_{Bj(SRE)}$ is the between-imputation variance for the mean residuals from the SRE model. Variances were calculated using the final replicate weights and appropriate replication method for the country.

3.2.2 Smoothed Variances

Since the direct or SRE estimates of the variances are subject to substantial sampling error, the true variances (or relative variances $\varphi_j^2 = \sigma_j^2/p_j^2$) were predicted using a modeling approach. A

requirement of this modeling is that the predicted variances should not depend directly on the SA-level SRE (or direct) estimates or variance estimates. An important feature of the development of the model for predicting the variances is that approximate values will suffice since the values of the relative variances affect the indirect estimates in only a minor way. Their main impact is in stabilizing the widths of the credible intervals.

The variance smoothing process was only done for proportions. The variance smoothing process for variances of averages takes on a different form, and therefore direct variances were used in the models for this initial research while finalizing the recommendation for the smoothing process (more below). In addition, the variance smoothing step was not performed for Sweden because the sample sizes for SAs were adequate. With the exception of Germany, the SRE estimates served as input to the variance smoothing process, as described below. For Germany, there were no SRE estimates, so the direct estimates served as inputs.

Since the relative variance of an SA estimate depends on the value of the SA's proportion at or below Level 1 in literacy, a two-step approach was implemented to produce model-dependent estimates of the relative variances. The approach followed the one implemented in the 2003 National Assessment of Adult Literacy (NAAL) SAE program (Mohadjer, Kalton, Krenzke, Liu, Van de Kerckhove, Li, Sherman, Dillman, Rao, and White, 2009; Mohadjer, Rao, Liu, Krenzke, and Van de Kerckhove, 2011). In step 1, the proportions at or below Level 1 in literacy were predicted from a simple regression model relating the SRE estimates $\hat{p}_j^{surv}$ to predictor variables. In step 2, the resulting predicted proportions from step 1 were used in a generalized variance function (GVF) model to smooth the SRE relative variance estimates.

For step 1, the logit of the SRE proportion of the population at or below Level 1 in literacy was used as the dependent variable in the regression model. A robust regression M-estimation approach using SAS Proc RobustReg was used to arrive at the predicted values of the proportion at or below Level 1 of literacy. Each SA was assigned a weight of the square root of its sample size on the grounds that it's sampling error—which was related to its sample size—was an important part of its residual error in the regression model. The square root was applied as an ad hoc method of approximating weighting by residual variance. The model had the form:

$$logit\left(\hat{p}_j^{surv}\right) = \gamma_0 + \gamma Z_j + e_j,$$

where

$\hat{p}_j^{surv}$ = the proportion of adults at or below Level 1 in literacy from the SRE model;

Z_j = the predictor variables (given in Table 3-3); and

e_j = the error term.

Table 3-3 Covariates used in variance smoothing

Country	Covariates						
Germany[1]	Gender (1 level)	Age (3 levels)	Nationality (1 level)	Education attainment (3 levels)	Employment status (2 levels)		
Italy	Gender (1 level)	Age (mean)	Citizenship (1 level)	Education attainment (2 levels)	Employment status (2 levels)	Number of people in the household (2 levels)	Marital status (1 level)
New Zealand	Gender (1 level)	Age (4 levels)	Birthplace (1 level)	Education attainment (4 levels)	Employment status (1 level)	Ethnicity (2 levels)	
Slovakia	Gender (1 level)	Age (4 levels)	Nationality (1 level)	Education attainment (3 levels)	Employment status (2 levels)	Language (1 level)	

Note: No variance smoothing was done for Sweden because no area-level SAE models were being fit.

In step 2, the predicted values of the proportions from the above regression model were used as predictor variables in the model to smooth the relative variance estimates. To make the model linear in the parameters, a robust weighted least squares log-log model was used, where the weight was the square root of the degrees of freedom for the direct variance estimate. The less precise relative variances have less impact in this ad hoc weighting scheme. The degrees of freedom was computed as *nunit* – 1, where *nunit* is the number of variance units within the SA. For Germany, the number of variance units within an SA was unknown, so it was assumed equal to 2 for all SAs. The robust regression approach was the same as the approach used in step 1. The model had the form:

$$\log\left(\varphi_{j(SRE)}^2\right) = \eta_0 + \eta_1 \log\left(\tilde{p}_j\right) + \eta_2 \log\left(1 - \tilde{p}_j\right) + \eta_3 \log\left(n_j\right) + e_j,$$

where

$\varphi^2_{j(SRE)}$ = the SRE relative variance of the proportion at or below Level 1 in literacy;

$\tilde{p}_j$ = the predicted proportion from step 1;

n_j = the sample size; and

e_j = the error term.

The predicted values of the relative variances for the SA proportions of adults at or below Level 1 in literacy were then computed based on the above GVF regression model, and these predicted values were treated as known relative variances in the small area models.

Smoothing variances for averages

As mentioned above, smoothing variances for averages was not conducted. However, since the computation of the results, we have developed a useful approach as follows. The purpose is to smooth the residual variance (from SRE). The variance of the average is smoothed by fitting a weighted least square model as below:

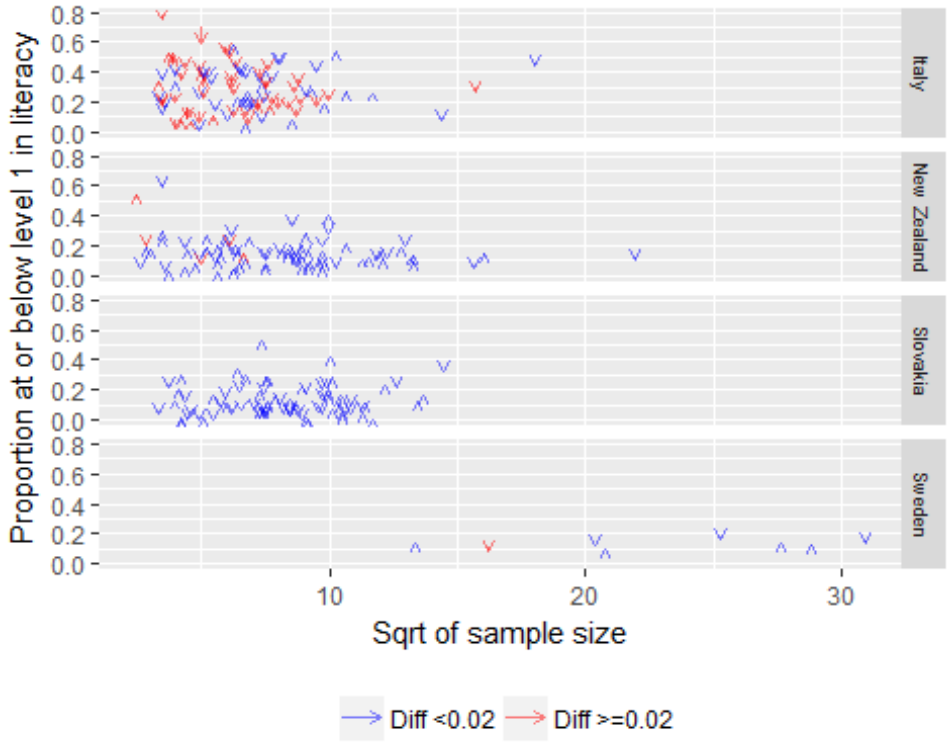
$$ln(Var_{r_a}) = \beta_0 + \beta_1 ln(C_a) + \beta_2 ln(B_a) + \beta_3 ln(\hat{\sigma}^2_{y_a}) + \epsilon$$

where $ln(Var_{r_a})$ is the natural log of the residual variance for each small area a, $ln(C_a)$ is the natural log of the number of clusters in each small area, $ln(B_a)$ is natural log of the average cluster size for each small area a, and $ln(\hat{\sigma}^2_{y_a})$ is the natural log of the estimated population variance of the proficiency scores among each small area a. The model is weighted by C_a. The exponentiation of the predicted value from this model is the smoothed variance. In this approach, the plausible value results can be combined first through the multiple imputation formula, and then smooth the combined variance, or it can be smoothed separately for each PV to continue the parallel PV processing (refer to Section 4.6).

3.3 Results

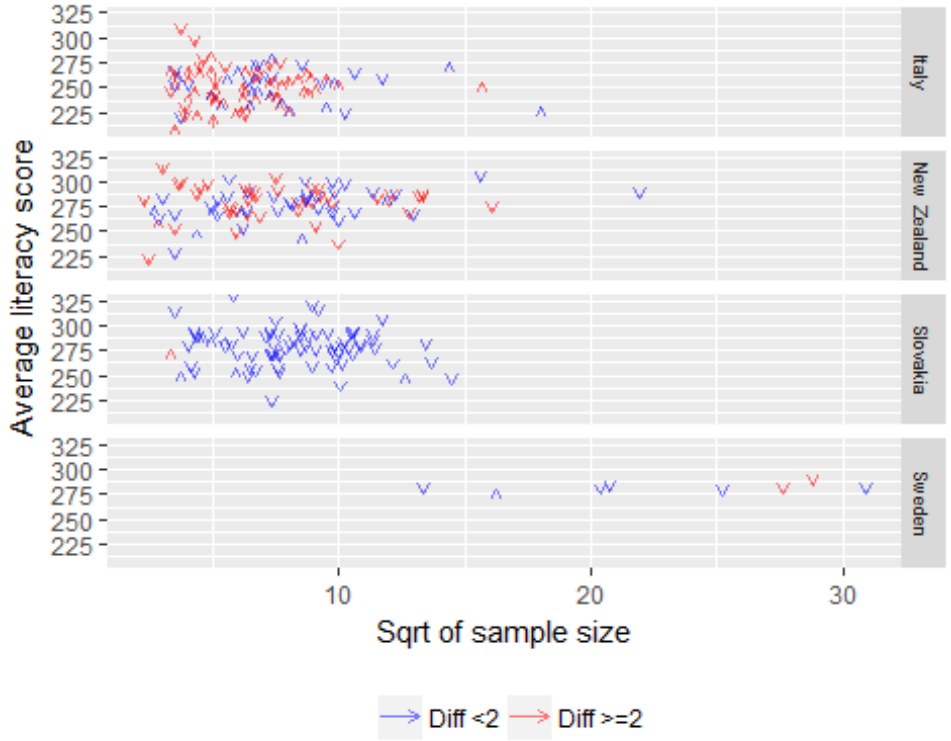
Figures 3-1 and 3-2 compare direct estimates and SRE estimates. Figure 3-1 is for estimates of the proportion at or below Level 1 in literacy, and Figure 3-2 corresponds to average literacy. The results for each country are shown in a shrinkage plot, with the arrow starting from the direct estimate and ending at SRE estimate. The x-axis is the square root of the sample size. Estimates that changed by more than 0.02 (or 2 percentage points) for proportions or by more than 2 for averages are highlighted as red. The results in Figure 3-1 for proportions indicate that the SRE had the largest impact on the point estimates for Italy, and it had the least effect on the point estimates for Slovakia. This could be related to the number and strength of available covariates (see Table 3-2). The results in Figure 3-2 for averages look similarly, especially for Italy. Diagnostic checks and further investigation may improve the fit of the models for averages.

Figure 3-1 Shrinkage plots comparing the direct and SRE estimates of the proportion at or below Level 1 in literacy



NOTE: No SRE was produced for Germany because Germany could not provide respondent-level data.

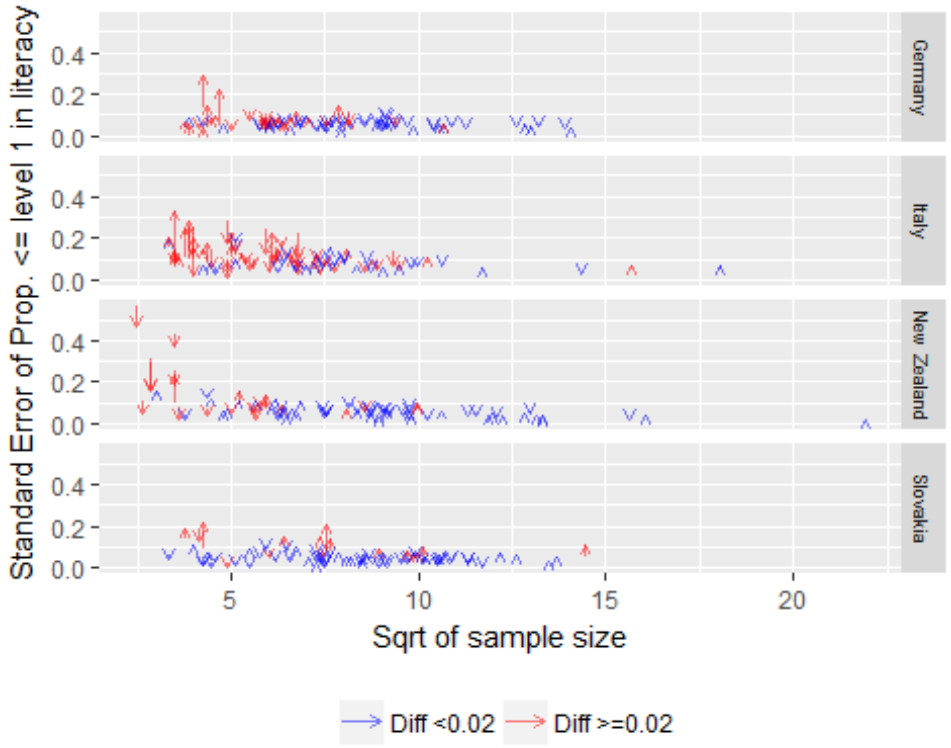
Figure 3-2 Shrinkage plots comparing the direct and SRE estimates of average literacy



NOTE: No SRE was produced for Germany because Germany could not provide respondent-level data.

Figures 3-3 shows the shrinkage plots comparing the direct and smoothed standard error estimates for the proportion at or below Level 1 in literacy. The smoothing process had a larger impact in the SAs with smaller sample sizes. This is expected, as the direct variance estimates are less stable in such SAs.

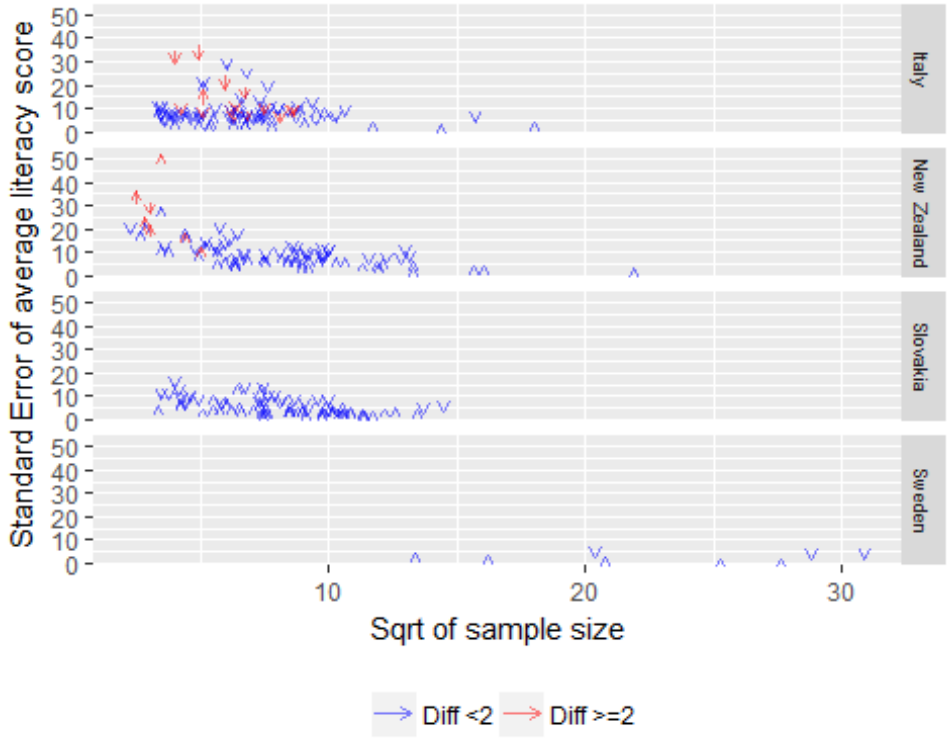
Figure 3-3 Shrinkage plots comparing the direct and smoothed standard error estimates for the proportion at or below Level 1 in literacy



Note: No variance smoothing was done for Sweden because no area-level SAE models were being fit.

As explained in Section 3.2, an SRE model was fit for average literacy but no additional variance smoothing was performed. Figure 3-4 shows the shrinkage plots comparing the direct and SRE standard error estimates for average literacy. As with the point estimates, the SRE had the largest impact on the standard errors for Italy, and little impact on those for Slovakia or Sweden. The effect is smaller for Sweden because the initial variance estimates were based on larger sample sizes and therefore more stable.

Figure 3-4 Shrinkage plots comparing the direct and SRE standard error estimates for average literacy



NOTE: No SRE was produced for Germany because Germany could not provide respondent-level data.

4. Models

As mentioned previously, models need to account for the variance impact from complex samples, which includes differential weighting in direct estimates, and clustering. If the sample of the small areas is not selected as a simple random sample, the sample design is informative. Also, weighting adjustments for nonresponse can reduce bias to the extent that the weighting variables are related to the proficiency scores. Being able to handle informative sampling needs to be considered and addressed because otherwise, the process could potentially lead to biased estimates.

In addition, the SAE models have to account for various sources of error and address them in the modeling approach. The various sources of error include the following.

- Sampling error results from probability sampling and the fact that different results would occur for repeated samples. Sampling error is addressed through the sampling model in the area-level model framework. Variances that result due to sampling error are smoothed and assumed known in the HB model.

- Model error results from estimation of model parameters, such as area-level random effects. This type of error accounts for different results occurring for different runs of the modeling process due to its random mechanism in fitting the models. Hierarchical Bayes (HB) methods typically account for the noise contributions attributed to estimating model parameters (beta coefficients). Model error is addressed through the use of area-level random effects, and accounting for noise when estimating model parameters (beta coefficients).

- Measurement error occurs when using covariates that are subject to sampling or nonsampling error. The covariates may have inaccurate measurements with possible systematic bias. Special attention to measurement error in the PIAAC model may be needed if covariates come from a survey with standard errors that are too large to ignore. Unit-level models are susceptible to measurement error when the definitions for covariates from external sources are not quite the same as for the survey. The choice of an area-level model helps in minimizing the effect of the measurement error that exists in covariates.

- Prediction error results from making estimates from the final model for areas without sample cases. Replications of the prediction process will achieve different results each time due to random draws relating to the random effects. Prediction error is addressed through the modeling process.

- Imputation error results from the generation of plausible values (PVs) and that different results would occur for replications of the imputation process. The PVs themselves

come from a model, and that uncertainty needs to be accounted for in the SAE estimation process.

- Model misspecification can occur when parameters estimated at the national level are used at the small area level. For example, if the small area is distinctively different, then bias will occur when the national parameter is applied. This occurs in PIAAC for example when the PVs are generated for national purposes, but are used for direct small area estimates. With the goal of lowering the mean square error (MSE), there is a balance between using more stable parameter estimates and introducing some bias in the results. Misspecification error is addressed by allowing the models to be estimated for each PV and combined to arrive at the small area point estimates and precision estimates through the multiple imputation formulas. This addresses misspecification error because each run on each PV takes into account the area-level random effects, and therefore reduces the misspecification bias that may be introduced in the item response theory (IRT) model that is based on national parameters.

4.1 Types of Models

In an area-level model, direct estimates produced at the local area-level are the prime elements in the modeling process. One part of an area-level model is a "sampling model," where survey-weighted estimates are produced for the small-areas with sample-design based variance estimates. The other part is the "linking model" (or regression model), which is developed using predictors at the small-area-level and could include variables at higher levels also. One can also distinguish between "matched" and "unmatched" models, where the former has the survey weighted estimate directly as the dependent variable in the model regression, and in the latter case, a functional transformation (e.g., the logit function) provides the link to the predictors, that is, the regression model and sampling model do not blend together directly.

Unlike the area-level approach, the unit-level model is built at a much lower level such as individual persons or households. That is, a unit-level model uses covariates available at the person-level to generate person-level values, which are aggregated to compute statistics at the area-level. There is potential for smaller MSE and for producing estimates for a wide range of other subgroups of interest. There is no effort to generate sample-design unbiased estimates. The basic unit-level models ignore sample-design based variance estimates at this very low level.[5]

[5] Survey weights can be used in estimating the parameters of the models, making these parameter estimates design consistent. The totals and the variance estimates would be entirely model-based, however.

If the model is linear, either the area-level or the unit-level approach could be used for a PIAAC small-area program. In the nonlinear case (e.g., in estimating a small proportion), a full cross-tabulation of the covariates is needed at the small area level. The area-level approach is more design-based, since the basic building blocks are the sample (design-based) estimates at the targeted local level, as well as the sample-design based variance estimates at this level. The unit-level approach is more dependent on the validity of the model, as it disaggregates down to the lowest levels. Sampling weights can be used to estimate the parameters of the model, which can make this portion of the estimation process sample-design consistent.[6] Variance estimates are entirely model dependent. Extensions have included a random-effect term as an attempt to capture the between area variation (see Rao and Molina, 2015).

Operationally, the area-level approach works with a much simpler data set, with one record for each local area rather than one record for each household or person, and in that sense is easier to work with in practice. This is especially useful as the Bayesian methods require numerous iterations with the data set as an input in each iteration.

4.2 Models Used in this Research

The models evaluated in this research were:

- Fay-Herriot (F-H) area-level model
- Hierarchical Bayes (HB) area-level matched (linear) model
- HB area-level unmatched (nonlinear) model (used when estimating proportions only)
- Unit-level empirical best linear unbiased predictor (EBLUP)

The F-H model is an area-level EBLUP model. The HB area-level model allows error from the model parameters to contribute to the mean square error of the small area estimates. The model statement for each is given in this section. Model statements are given under the context of estimating the proportion at or below Level 1, or the average, and can be applied similarly to any literacy level.

[6] That is, the estimates are approximately unbiased estimators of the corresponding parameters at the population level, over all possible samples, with this property not dependent on the validity of the model.

F-H Area-level Model for Proportion at or Below Level 1 or Average Literacy

The Fay-Herriot model (Fay and Herriot, 1979) employs a sampling model and a linking model as follows:

$$\hat{y}_j = y_j + e_j \text{ (sampling model)}$$

$$y_j = x_j'\beta + u_j \text{ (linking model)}$$

where

$\hat{y}_j$ = direct estimator of y_j for area *j*;

e_j = sampling error;

u_j = area-specific random effect;

$e_j \sim$ $N(0,\sigma_e^2)$;

$u_j \sim$ $N(0,\sigma_u^2)$.

Combining the sampling and linking models leads to:

$$\hat{y}_j = x_j'\beta + u_j + e_j \qquad (4a)$$

The σ_e^2 are typically smoothed through the use of generalized variance functions and treated as if known.

Linear Matched Area-level Model for Proportion at or Below Level 1 or Average Literacy

The area-level linear matched model applies HB estimation as given in Rao and Molina (2015). The combined model (4a) is estimated under the following assumptions:

$e_j \sim N(0,\sigma_e^2)$;

$u_j \sim N(0,\sigma_u^2)$; and

flat prior distributions for $\beta, \sigma_e^2, \sigma_u^2$.

Unmatched HB Area-level Model for Proportion at or Below Level 1

The unmatched HB area-level model used in this research is a one-random effect version of the two-random effect model that was developed under contract to the National Center for Education Statistics (NCES) as applied to the 2003 National Assessment of Adult Literacy (NAAL) (see, for example, Mohadjer et al., 2011).[7] In the unmatched model case, as introduced by You and Rao (2002), there is an intermediate function between the sample estimate and the linear predictive model. The intermediate function (logit) may be necessary due to small sample size as well as the low estimated proportions. The NAAL small area program used an HB approach (see, for example, Rao and Molina (2015), Chapter 10) rather than the simpler Empirical Bayes (EB) approach because of the use of the logit link in the regression model.

As applied to this research, the direct estimator of p_j within small area *j* is $\hat{p}_j$, which is subject to both sampling and measurement error[8], combined into a single error term e_j. The dependent variable z_j in the mixed-effects regression model is the logit of the true proportion p_j (the logarithm of the odds, where the odds is the ratio $p_j/(1-p_j)$). The following sampling model and mixed effects regression linking model were used in this research:

$$\hat{p}_j = p_j + e_j$$

$$z_j = x_j'\beta + u_j$$

where $z_j = \ln(p_j/(1-p_j))$, and

where β are fixed regression parameters, v_j is an area-level random intercept representing the difference between the true value of the characteristic for the area and its model-based expectation.

The random effects u_j and e_j were assumed to be normally distributed with mean zero and variances σ_u^2, and σ_e^2, respectively, and were assumed to be independent. The HB approach used a flat prior distributions for β, and gamma priors for σ_e^2, σ_u^2.

[7] The process was repeated for the 1992 National Adult Literacy Survey (NALS).

[8] The measurement error is from the assessment.

Unit-Level Model for Proportion at or Below Level 1 or Average Literacy

The basic unit-level model used is the empirical best linear unbiased predictor (EBLUP), which is as follows:

$$\hat{y}_{jk} = \beta_0 + x'_{jk}\beta + u_j + e_{jk},$$

where

$\hat{y}_{jk}$ = average of the 10 plausible values;

x'_{jk} = covariates for the respondents k in small area j;

u_j = area-specific random effect; and

e_{jk} = sampling and measurement error.

The model was performed with the following assumptions:

$e_{jk} \sim N(0,\sigma_e^2)$; and

$u_j \sim N(0,\sigma_u^2)$.

Any clustering that exists within the small area from the selection of primary sampling units (PSUs), secondary sampling units (SSUs) and in households (where two persons are selected) is not taken into account in the above model. If the within-small-area clustering exists, other unit-level models may need to be investigated, such as a model introduced by Stukel and Rao (1999) called a two-fold nested error regression model, or bootstrapping. Otherwise it is best to pursue an area-level model.

For proportions, because the outcome is binary (at/below Level 1 or not), a generalized linear mixed model (GLMM) with an area-level random effect could have been considered for the proportion at or below Level 1.

$$g(\hat{p}_{jk}) = \beta_0 + x'_{jk}\beta + u_j + e_{jk},$$

where

$\hat{p}_{jk}$ = probability of being at or below Level 1;

x'_{jk} = covariates for the respondents k in area j;

u_j = county-specific random effect;

e_{jk} = measurement error; and

$g(\ )$ = logit link function.

When the true proportions are not at the extremes, say, approximately in the interval (0.2, 0.8), the logit link function is approximately linear. In this case, fitting a linear mixed model (LMM) is approximately equivalent to fitting the GLMM with logit link even if the outcome is binary. Molina and Strzalkowska-Kominiak (submitted for publication) showed that plug-in estimators of area proportions based on a unit-level GLMM with logit link performed similarly to the much simpler LMM in the example of estimating the true activity rates in Switzerland, which are not likely to be very extreme. Therefore, we have applied the LMM to proportion at or below Level 1 in this research. Table 4-1 provides the covariates used for each country in the models and their pairwise correlations with direct estimates. The correlations with the direct estimates for proportion at or below Level 1 ranged from -.33 to .53 for Slovakia, -.51 to .49 for New Zealand, and -.42 to .45 for Italy, and a bit lower in magnitude for Germany (-.11 to .07). The correlations with the direct estimates of averages ranged from -.52 to .29 for Slovakia, -.56 to .51 for New Zealand, and -.45 to .45 for Italy, and a bit lower in magnitude for Germany (-.12 to .18).

Table 4-1. Summary of Model Covariates and their pairwise correlations with direct estimates, by Country

Country	Area-level model covariates and correlations			Unit-level model covariates
	Covariates (levels)	Correlations with direct estimate of proportion	Correlation with direct estimate of average	
Germany	Gender (1) Age (3) Education (3) Employment (2) Nationality (1)	.06 -.10, -.06, .03 -.03, .07, -.10 -.11, .06 .01	-.12 .18, .15, -.06 .05, -.12, .17 .15, -.09 -.07	NA
Italy	Gender (1) Age (mean) Education (2) Employment (2) # of people in household (2) Citizenship (1) Marital status (1)	.02 -.32 .41, -.31 -.36, .27 -.42, .29 .25 .45	-.08 .37 -.39, .30 .41, -.33 .45, .25 -.31 -.45	Gender (1 level), Age (mean), Education attainment (2 levels), Employment status (1 level), Number of people in the household (2 levels)
New Zealand	Gender (1) Age (4) Education (4) Employment (1) Birthplace (1) Ethnicity (2)	.05 .27, .06, -.26, -.17 .44, -.01, -.41, -.36 -.47 .02 -.51, .49	-.03, -.27, -.04, .37, .14 -.56, -.06, .35, .51 .48, -.11 .48, -.54	Gender, Age groups (4 levels), Birthplace (1 level)
Slovakia	Gender (1) Age (4) Education (3) Employment (2) Language (1) Nationality (1)	-.32 .46, -.08, -.14, -.15 .53, -.02, -.36 -.33, .34 -.16 -.11	.24 -.39, .05, .07, .17 -.52, .08, .31 .29, -.34 .17 .15	Gender (1 level), Age groups (4 levels), Education attainment (2 levels)
Sweden	NA			Gender (1) Age (4) Birthplace (1)

Note: As pointed out in Lahiri and Suntornchost (2015) the true population correlations are higher. The correlation estimates could be improved if the sampling error is taken into account, as described in Lahiri and Suntornchost (2015).

4.3 Predictions

The prediction process is explained below for each model, first for areas with PIAAC sample, and second for areas without PIAAC sample.

Predictions for Areas with PIAAC Sample

For the area models, the final estimates at the small area level are combinations of model predictions and direct estimates for areas with PIAAC survey data. If there is sample in all small areas, the estimates are not synthetic, but rather take in at least some information from the PIAAC data

directly. For the F-H model, the best linear unbiased predictor (BLUP) is derived as a composite of the direct and model based estimate:

$$\tilde{y}_j = \alpha \hat{y}_j + (1 - \alpha) x_j' \tilde{\beta} \quad (4b)$$

Where,

$\alpha = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$

$\tilde{\beta}$ is a weighted least squared estimator of β.

For the HB area-level models, one run resulted in B = 10,000 MCMC samples where estimates of the model parameters were obtained. This was achieved after a burn-in of 10,000 iterations. Because the results from neighboring iterations after burn-in are correlated, they were "thinned" by taking a systematic sample of 1-in-10 of them. The HB estimates are computed as:

$$\tilde{y}_j^{HB} = (1/B) \sum_{b=1}^{B} \tilde{y}_j^{(b)},$$

where, $\tilde{y}_j^{(b)}$ is obtained from the b^{th} MCMC sample using the combined formula (4a) for $\tilde{y}_j$.

For the unit-level model, the EBLUP predictions are a direct result from applying the estimated model parameters to the covariate data.

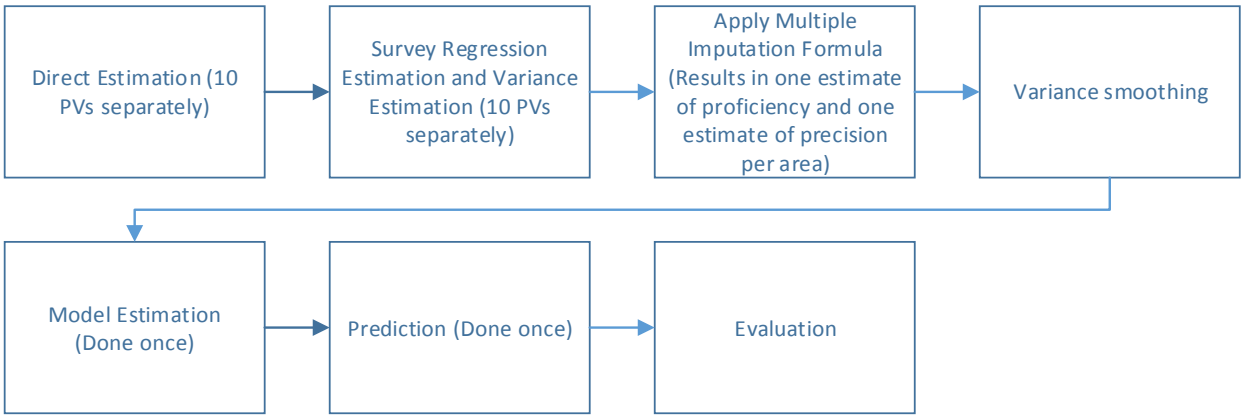
Predictions for Areas without PIAAC Sample

The final estimates at the small area level are entirely model predictions for areas without PIAAC survey data. For the F-H model, only the second term in (4b) is used. For HB area-level models, the b^{th} MCMC sample estimate of the area-level random effect is not available. In this case, a random draw is taken from the normal distribution with mean zero and corresponding variance for the b^{th} MCMC sample. For the unit-level model, the EBLUP predictions are a direct result of applying the estimated model parameters to the covariate data.

4.4 Addressing Imputation Error Variance

For the initial research, the handling of PVs was conducted as illustrated in Figure 4-1. As shown in the figure, each step prior to the model estimation was repeated for each PV. Then the multiple imputation formula, as shown in Rubin (1987), was applied to combine the results into one estimate for each small area, with one variance estimate, which has contributions from sampling error and imputation error components. The model is estimated once from the post-combined estimates. Hence, the prediction phase and evaluation is also done once. This process assumes the variance contribution from PVs is known when estimating the beta coefficients. An improvement to this process is discussed in Section 4.6.

Figure 4-1 Addressing imputation error in the initial research



4.5 Estimates of Precision

Mean square errors (MSEs) were computed for all estimates derived from the model estimation and the prediction process for all models used in this research. For the HB area-level models, credible intervals were also computed as the 2.5th and the 97.5th percentiles of $\tilde{y}_{ij}^{(b)}$, where b = 1, …, B. With the Phase 2 recommended approach (discussed below) of parallel PV processing, the credible intervals would need to be adjusted under the HB model to account for the imputation error variance component.

4.6 Improvements to the Research Models

This section discusses some extensions to the research models that may lead toward improvements for publishable estimates.

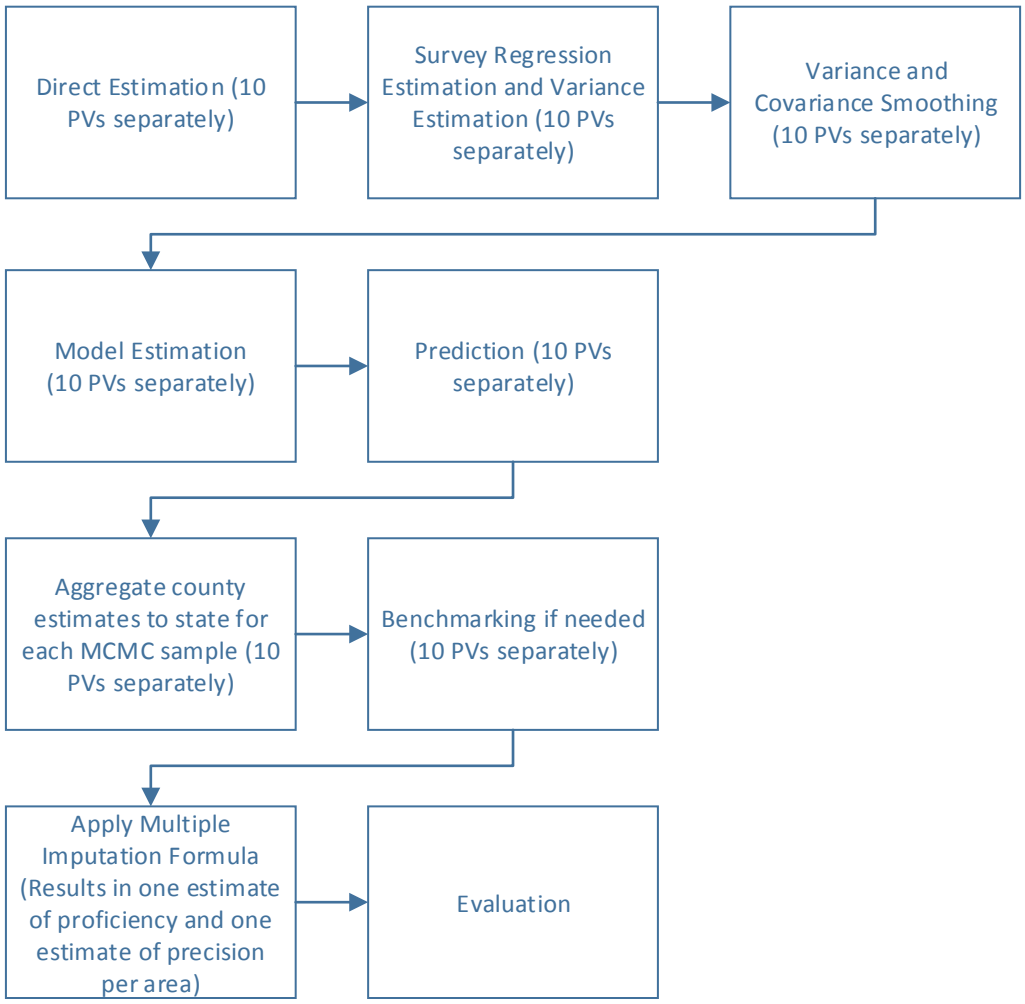
Parallel PV Processing

Imputation error could possibly be addressed more fully by repeating the entire direct estimation, smoothing, model estimation, prediction process and benchmarking (if conducted) for each PV. Then the results would be combined through the traditional multiple imputation formula. Figure 4-2 is a flowchart of the parallel PV process. This is an on-going research, being conducted for the U.S. sample, to see whether this process addresses the issue of assuming the variance contribution from PVs is known when estimating the beta coefficients.

Multivariate Models

For PIAAC, small area estimates can be considered for multiple statistics, for example, proportions for Level 1 or below, Level 2, and Level 3 and above, as well as averages, for the two domains: literacy and numeracy. The models discussed above could be processed in separate univariate models, perhaps using the same covariates, in lieu of a multivariate model. With separate univariate models, two of the three levels would be modeled and the third level would be derived from the model results of the first two levels because the sum of the three proportions would add to one. Under contract to NCES, Westat has conducted research in a multivariate HB linear matched model that will result in estimates for Level 1 or below, Level 2, and Level 3 and above. The model takes advantage of the covariance between domains, which may result in reduced MSE. Due to the demands on the model fitting and if there is a small number of data points, it may be best to fit the multivariate model for proportion separately for literacy and numeracy.

Figure 4-2 Processing the 10 plausible values to address for model misspecification and imputation variance



Multiple Random Effects

The above models use the small area-level random effect, but it may be of interest to use a two- or three- levels of random effects. For example, an extension to the Fay-Herriot model (Torabi and Rao, 2014) would employ a sampling model and a linking model as follows:

$$\hat{y}_{ij} = y_{ij} + e_{ij} \text{ (sampling model)}$$

$$y_{ij} = x'_{ij}\beta + v_i + u_{ij} \text{ (linking model)}$$

where

$\hat{y}_{ij}$= direct estimator of y_{ij};

e_{ij}= sampling error;

u_{ij}= area-specific random effect;

v_i = specific random effect for larger areas;

$e_{ij}\sim$ N(0,σ_e^2);

$v_i\sim$ N(0,σ_v^2);

$u_{ij}\sim$ N(0,σ_u^2).

Combining the sampling and linking models leads to:

$$\hat{y}_{ij} = x'_{ij}\beta + v_i + u_{ij} + e_{ij}$$

The σ_{ij}^2 are typically smoothed through the use of generalized variance functions and treated as if known. The best linear unbiased predictor (BLUP) for the sampled small areas is then derived as shown in equation 2.10 of Torabi and Rao (2014), and given here:

$$\tilde{y}_{ij} = x'_{ij}\tilde{\beta} + \tilde{v}_i + \tilde{u}_{ij},$$

Torabi and Rao (2014) provide the estimation details for the parameters, specifically for $\tilde{\beta}$ on page 38 of the reference (equation 2.6), for $\tilde{v}_i$ in equation 2.8 and for $\tilde{u}_{ij}$ in equation 2.9.

For non-sampled small areas, the pseudo-BLUP estimator is used as shown here following equation 2.11 of the reference:

$\tilde{y}_{ij}^* = x'_{ij}\tilde{\beta} + \tilde{v}_i$. See other details as given in Torabi and Rao (2014) near equation 2.11.

Another example of using two random effects is the NAAL model (see Mohadjer, et al 2011). The benefits of using two or more levels of random effects in the model are 1) benchmarking the

estimates may not be necessary as estimates are controlled through the random effects, 2) estimates for the small areas without sample will not be synthetic if all larger areas have PIAAC sample, and 3) associations of small areas within larger areas may have some impact while the same random effect applied to those areas.

Multivariate Two-fold Area-level HB Matched Model

A two random effect multivariate HB matched model may employ the traditional small area estimation structure, including a sampling model and a linking model, using matrix form notation to account for multiple subgroups (e.g., at or below Level 1, Level 2, Level 3 and above) as follows:

$$\hat{\boldsymbol{y}}_{ij} = \boldsymbol{y}_{ij} + \boldsymbol{e}_{ij} \text{ (sampling model)}$$

$$\boldsymbol{y}_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{v}_i + \boldsymbol{u}_{ij} \text{ (linking model)}$$

where

$\hat{\boldsymbol{y}}_{ij}$= direct estimator of $\boldsymbol{y}_{ij}$;

$\boldsymbol{e}_{ij}$= sampling error;

$\boldsymbol{v}_i$ = area-specific random effect for larger area *i*;

$\boldsymbol{u}_{ij}$= area-specific random effect for small area *j*;

$\boldsymbol{e}_{ij} \sim N(0,\boldsymbol{\Sigma}_{ij})$;

$\boldsymbol{v}_i \sim N(0,\boldsymbol{\Sigma}_v)$;

$\boldsymbol{u}_{ij} \sim N(0,\boldsymbol{\Sigma}_u)$.

Combining the sampling and linking models leads to:

$$\hat{\boldsymbol{y}}_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{v}_i + \boldsymbol{u}_{ij} + \boldsymbol{e}_{ij}$$

5. Model Diagnostics and Evaluation

Large-scale small-area estimation (SAE) programs generally employ an extensive model diagnostics and evaluation process because models are never a perfect fit to the data, and systematic errors can manifest themselves. The smaller the proportion of small areas (SAs) with PIAAC sample, the more important the evaluation becomes. Diagnostics are needed to ensure the model fit, and the evaluation is needed to ensure the results make sense and that the process performed as expected. Section 5.1 first discusses, for this initial research, a limited number of methods that were used to evaluate the fit of the SAE models to the direct estimates. The evaluation tools were selected to be able to identify appreciable problems with the initial research models. Section 5.2 provide the evaluation results. Section 5.3 discusses a number of approaches available for evaluating the SAE models and to be included in Phase 2.

5.1 Evaluation for the Initial Research

The evaluation conducted in this initial research is mainly graphical, as influenced by Khan (2018). The graphs were used initially, and if more investigation was warranted, the data itself were reviewed to better understand the issue at hand.

- Histograms of difference from direct estimates for each country. The main objective of this plot is to take a first look at the results through reviewing the distribution of the differences. The graph can also indicate outliers that would need more investigation, especially to check the size of the sample in those areas.

- Bubble plots of direct estimates by each model result, with size of bubble related to sample size. One would hope to see from this plot the large bubbles along the diagonal line, assuming that we would trust more in the direct estimates for larger sample sizes. With that assumption, if there are any outliers, they should be the small bubbles.

- Shrinkage plots with arrows showing the direction from direct estimates to model estimates, by sample size. The main purpose of this plot is to show how the model impacts the estimates. There should be some shrinkage, that is, the estimates are pulled toward the average, if the estimates are more dependent on the model than the direct estimates. The longer arrows show larger impact from the model, which should occur for areas with smaller sample sizes than others.

- Interval coverage plots, showing the confidence or credible interval of the model estimate, and the direct point estimates, by sample size. These plots help show whether

or not the resulting confidence/credible interval from the model covers the direct estimate. Again, the focus is solely on the areas with the largest samples. If the interval does not cover the direct estimates in the areas with largest sample sizes, it may indicate that the modeling process can be improved.

- MSE plots showing the resulting MSEs from direct estimation and models, by sample size. While the aforementioned graphs are used to review the point estimates, this plot shows the impact on precision. This review is a common aspect of SAE evaluations, and as an example, Bijlsma et al. (2017) also reviewed the decrease in standard errors from the SAE approach. Erciulescu et al. (2017) reviewed the outcomes to ensure (1) the negative correlation between coefficient of variance (CV) and sample size, and (2) the impact of the SAE approach on the CV. Mohadjer et al. (2011) reviewed the credible interval widths and CVs (direct and indirect) for sampled counties and nonsampled counties to ensure the impact of direct estimates on the indirect estimates' MSE for counties with sample. In general, the MSEs associated with the SAEs should be smaller than the MSEs associated with the direct estimates. If not, it may be due to weak covariates used in the models. Table 4-1 shows a measure of strength of the covariates used in the SAE models for each country.

- A table showing a comparison of the national direct estimate and the SAE estimates aggregated to the nation, weighted by population size of the small areas.

5.2 Evaluation Results

The results from the evaluation are provided below, first for the proportion at or below Level 1, and next for average literacy.

Special Note on Sweden

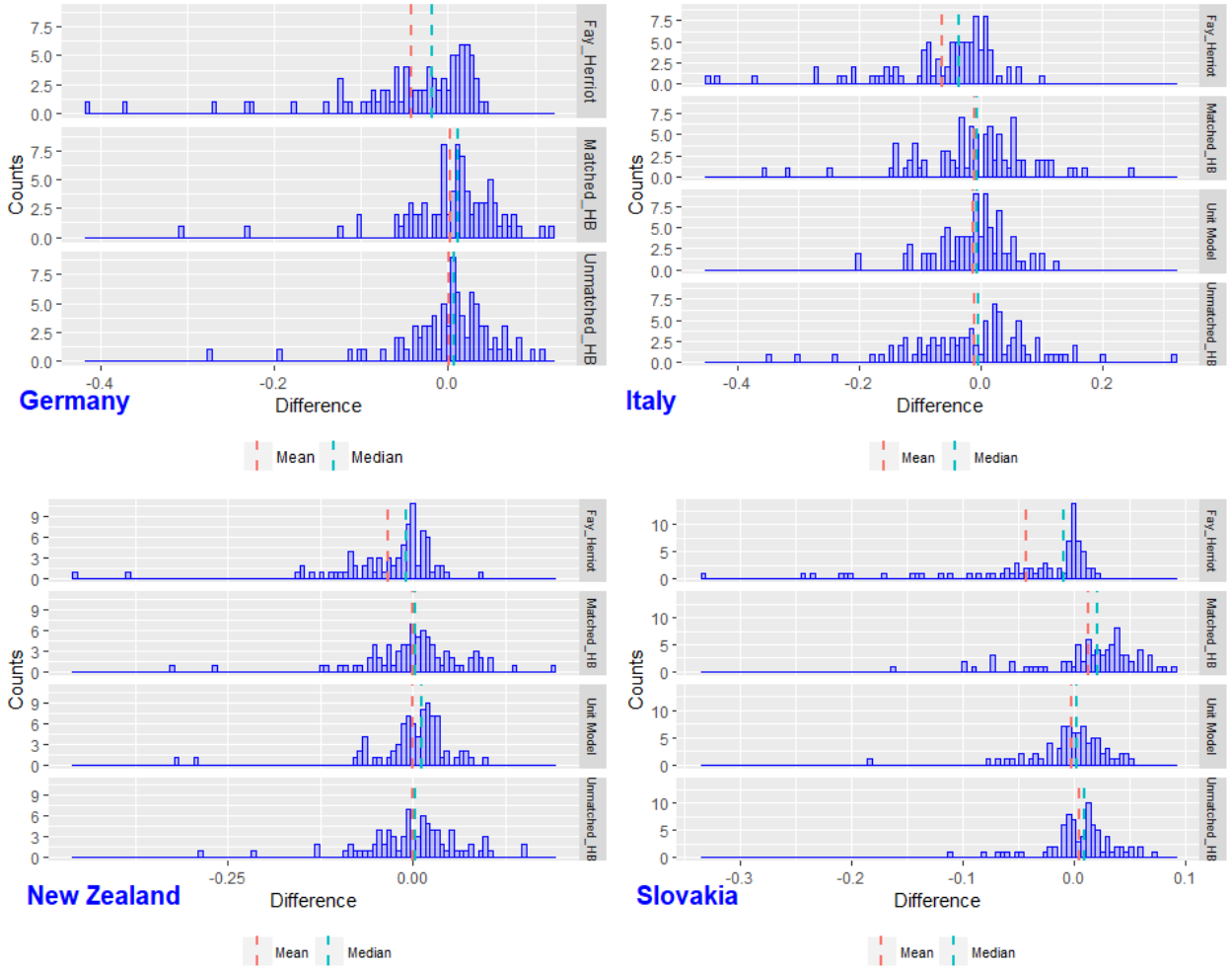
Because area-level models were not processed for Sweden due to the low number of SAs, they are excluded from the discussion below. Results of the application of a unit-level EBLUP SAE model is given in Appendix B. In general, Appendix B displays a demonstration of the unit-level EBLUP as applied to Sweden's data. The unit-level EBLUP does not take the PIAAC survey weights or any sample design features into account. Instead, it assumes that the units are selected from a simple random sample design. The covariates of the model are limited to indicators derived from age, sex and native born status. The education level is usually a stronger predictor but it is not available for use in Sweden's case. The survey weighted direct estimator and the SRE estimator are considered more reliable because the sample sizes in each of the eight small areas are fairly large to produce estimates of acceptable quality. The unit-level EBLUP can be improved in the future by

incorporating the weights and clustering (for countries with clustered designs) into the model specification.

Proportion At or Below Level 1

The differences between model estimates and the direct estimates are shown in the histograms in Figure 5-1 for all the models that were fitted to data from the four countries: Germany, Italy, New Zealand and Slovakia. The means and medians of the histograms are around zero except that the model estimates based on the Fay-Herriot model results are slightly smaller than the direct estimates on average. Majority of the differences are within 10 percentage points. Slovakia has fewer outlying differences as compared to the other countries. The outliers in the plots show that for a few small areas the model estimates can deviate from the direct estimates by about 20 to 40 percentage points.

Figure 5-1 Proportion at or below Level 1: Histogram of differences between SAE and direct estimate

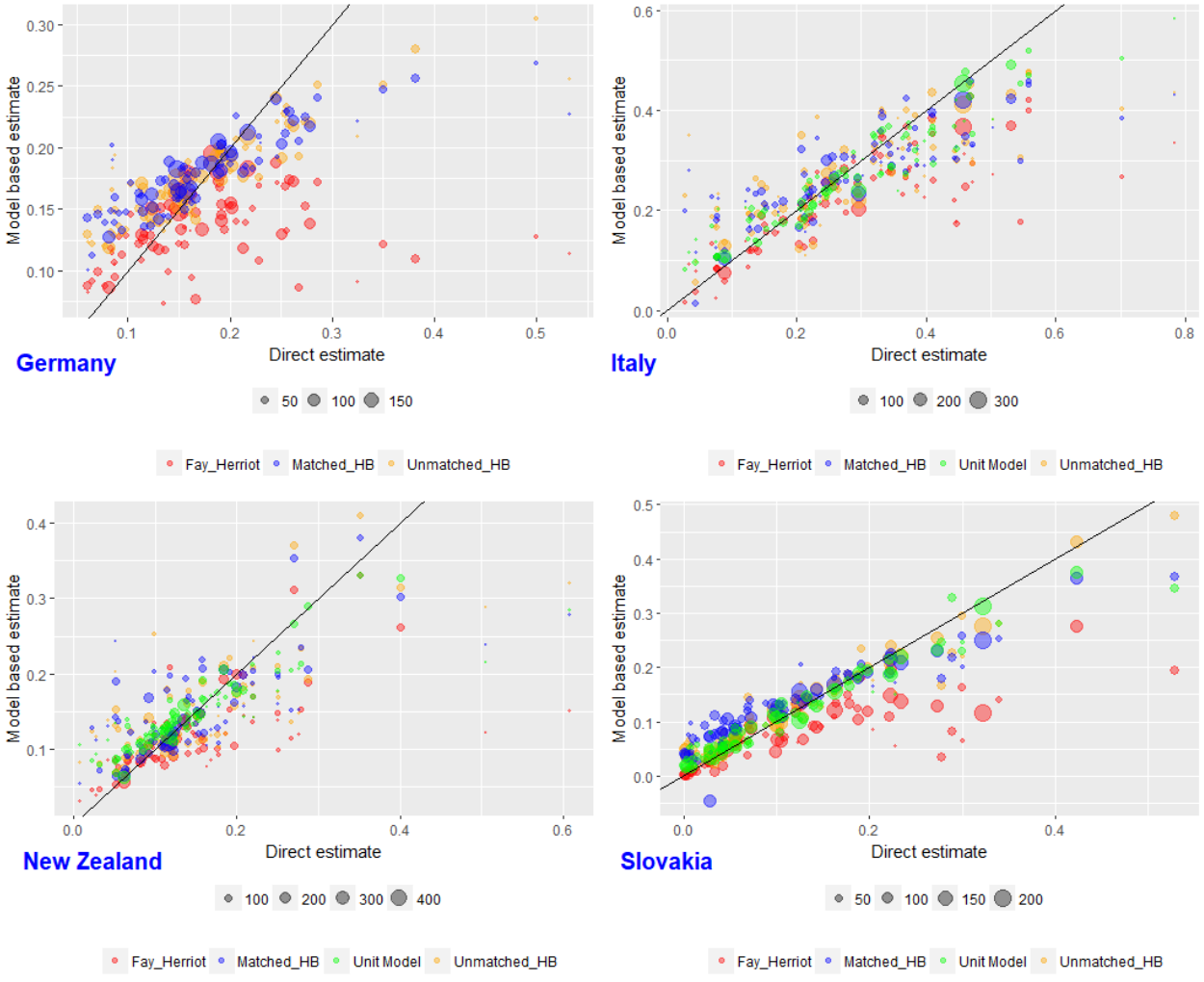


In Figure 5-2, most of the bubbles are located around the 45 degree lines, indicating that the direct estimates and the model estimates are close to each other. Some of the small bubbles, with the sizes of bubbles being proportional to the sample sizes in the small areas, are farther away from the 45 degree lines. This is as expected because the direct estimates contribute less to the model estimates when derived from samples of smaller sizes and associated with higher sampling errors (i.e., less reliable). The bubbles in the plot show that the model estimates are usually smaller than the direct estimates when the estimated proportions are larger than 20 percent, with the Fay-Herriot results being more extreme than the other models. Some investigation into Germany's results has shown that there is at least one very small smoothed standard error that is influential to the Fay-Herriot results. Removal of the influential case provides results very close to the matched HB model results.

In Phase 2, the smoothing model would be further investigated to determine the way to address the influential outlier.

The area level covariates used in Germany's models are from the 2011 Micro Census but they are also estimates associated with sampling errors. In addition, weak associations are observed between the area level covariates and the direct estimates. As a result, the models have low predicting power and may not work well for improving the quality of the direct estimates.

Figure 5-2 Proportion at or below Level 1: Scatterplot of SAE and direct estimates, with sample size as bubbles

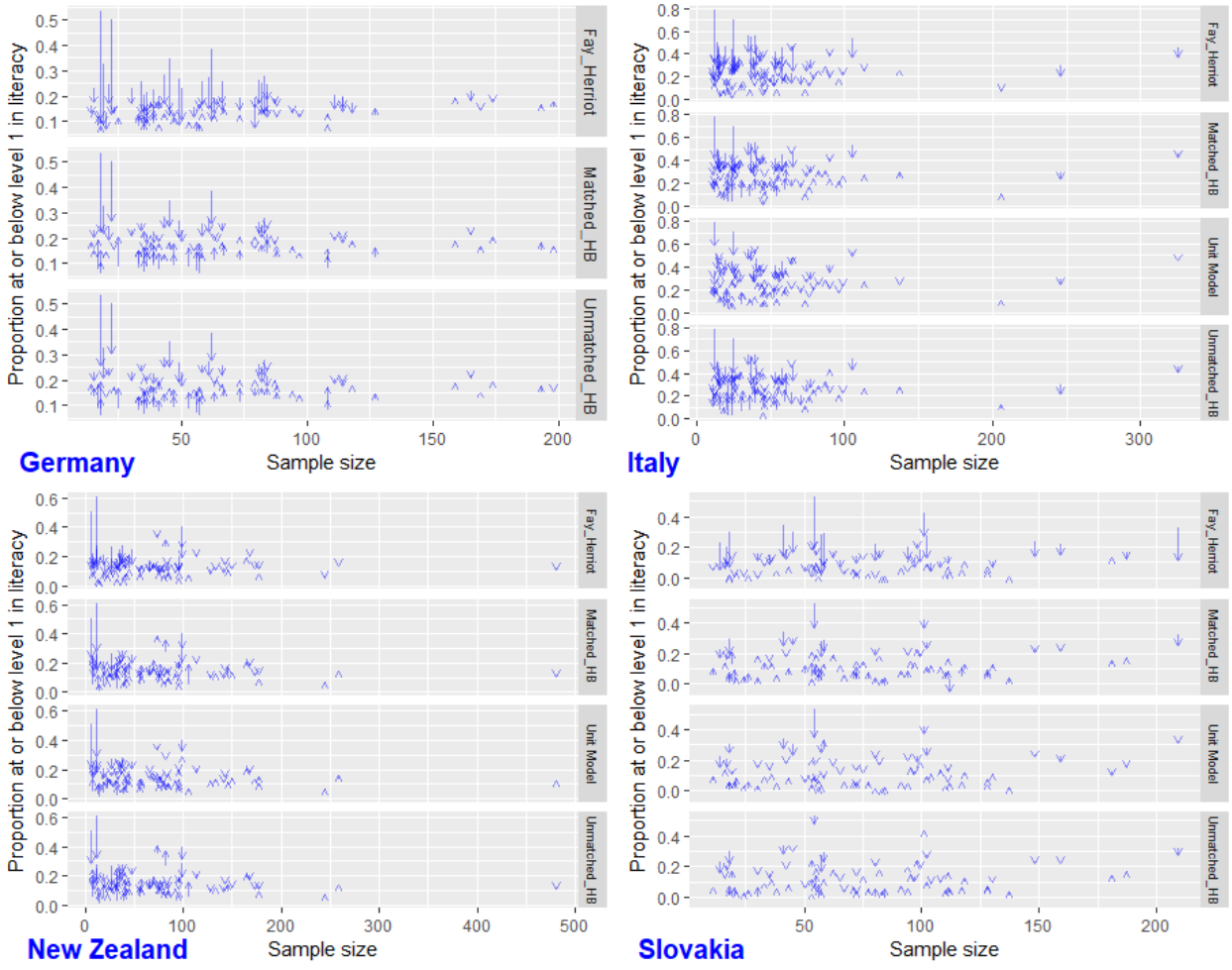


Shrinkage towards the means can be observed in Figure 5-3. The shrinkages are more significant, as expected, in areas with smaller sample sizes than those in areas with larger small sizes. Overall, less shrinkage is observed in Slovakia's estimates. The model estimates and the direct estimates become

much more similar when the sample sizes are above 100 to 150. There is a data point in Slovakia, where the Fay-Herriot result is much different from the direct estimate. This is an example of a result that would need further investigation. The estimates from the other models are close to the direct estimate for those two data points.

The models take advantage of the covariates that are highly predictable for the statistics of interest. When the covariates are highly correlated with the area estimates of interest, the models can be very helpful for generating reliable estimates compared to the direct estimates that may involve large sampling or imputation errors.

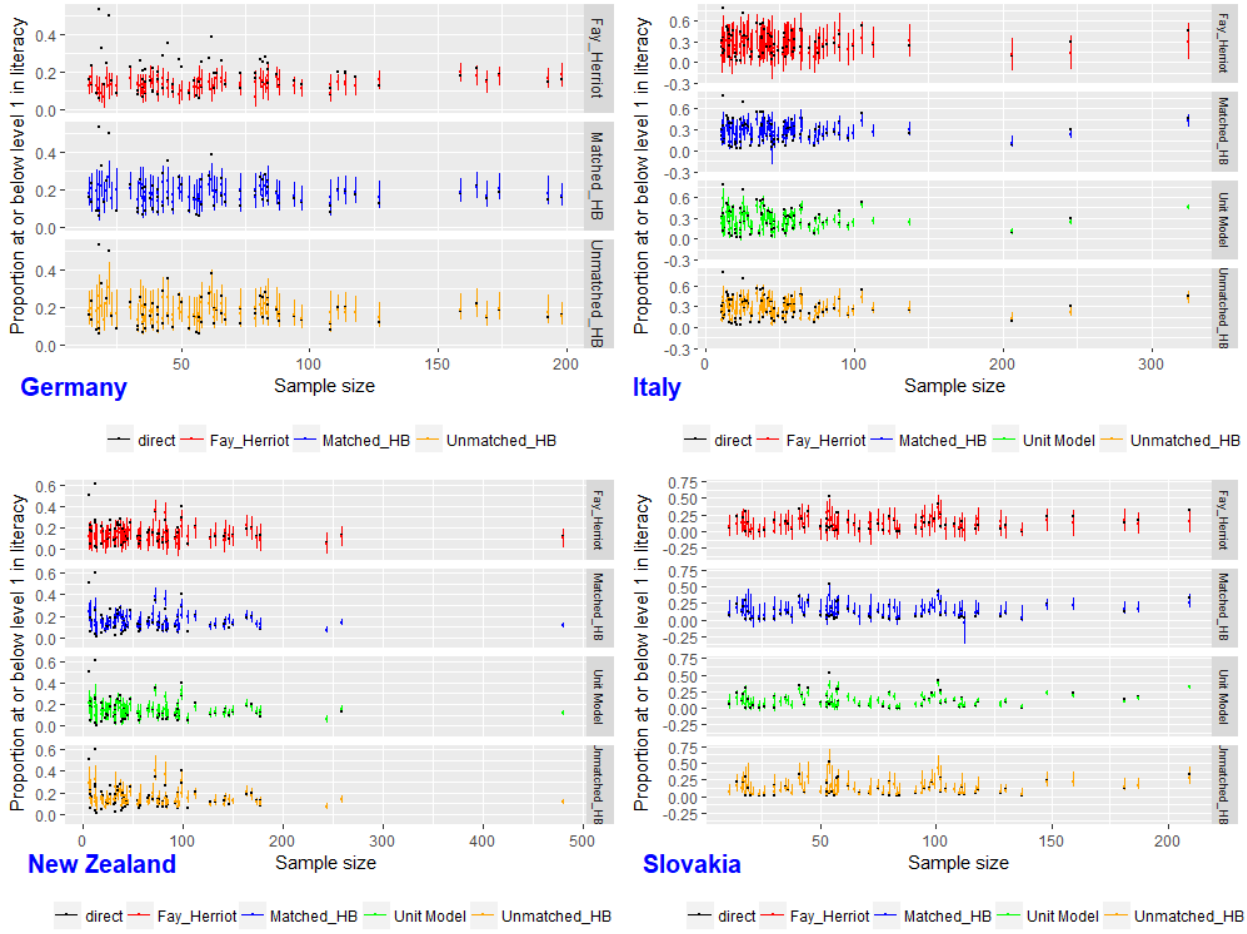
Figure 5-3 Proportion at or below Level 1: Shrinkage plots of point estimates by sample size



The interval coverage plots in Figure 5-4 show that, for majority of the small areas, the confidence/credible intervals generated from the models cover the direct estimates, especially for

areas with large sample sizes. When the sample sizes are less than 30, sometimes the confidence intervals from the models fail to cover the direct estimates. This is exemplified most notably in the plot for Germany, where several direct estimates are above the upper bounds of the confidence intervals from the Fay-Herriot model when the sample size is smaller than 100. In general, the predicting power of the linear model being weak due to the low correlation between the covariates and the direct estimates, which contributes to the width of the confidence interval.

Figure 5-4 Proportion at or below Level 1: Indication of coverage by confidence/credible interval



Figures 5-5 shows the direct estimates and all model estimates in one plot for each country. For Germany, the Fay-Herriot estimates are smaller in general than the other model estimates and many direct estimates. For Slovakia, the Fay-Herriot model and the matched HB model produce a couple of negative estimates. Both models assume linear relationship between the proportions and the

covariates and in these areas the direct estimates are almost as low as zero. In this case the model estimates can be truncated at zero or at a small proportion such as 0.01.

Figure 5-5 Proportion at or below Level 1: Comparison of point estimates between direct and SAE approaches

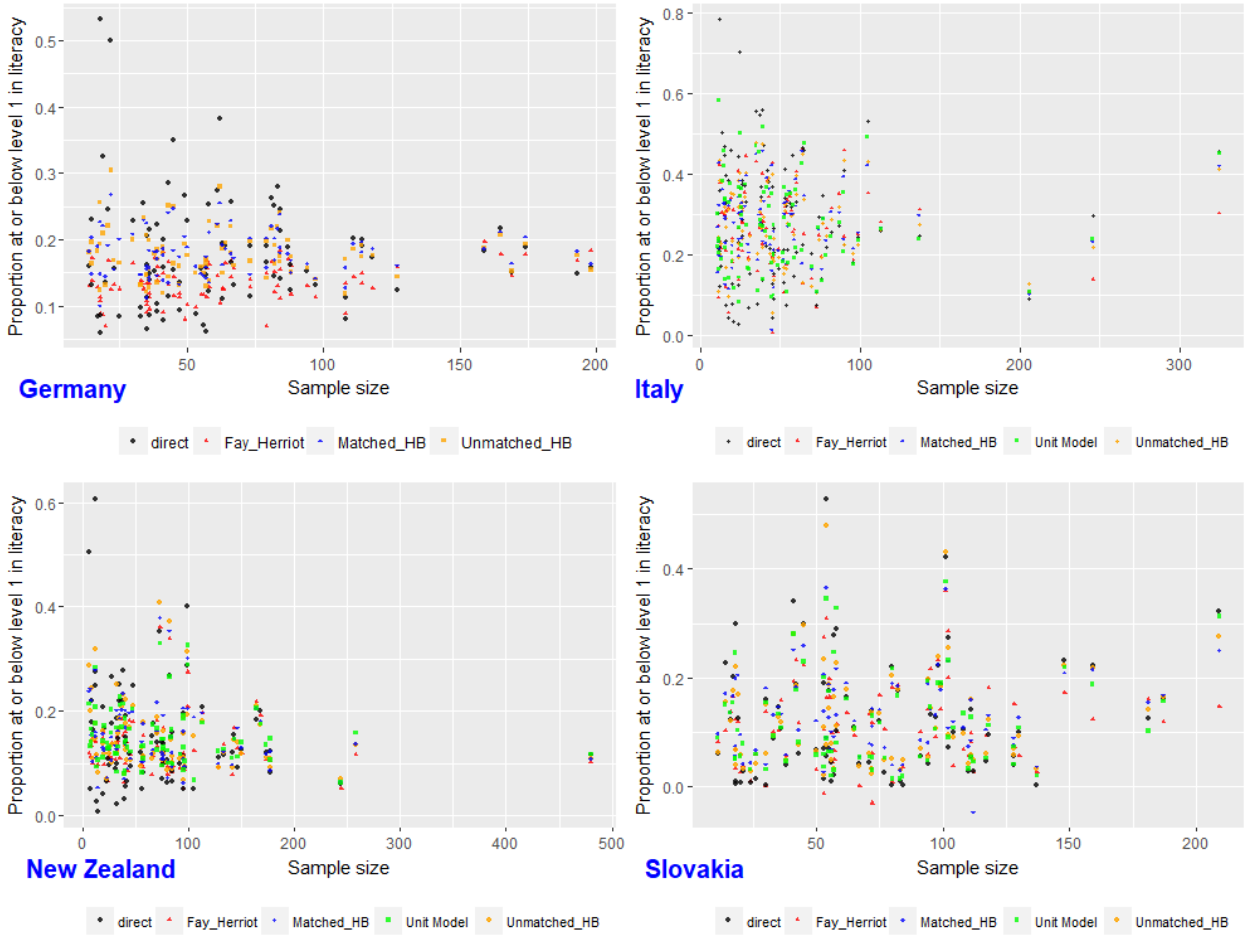
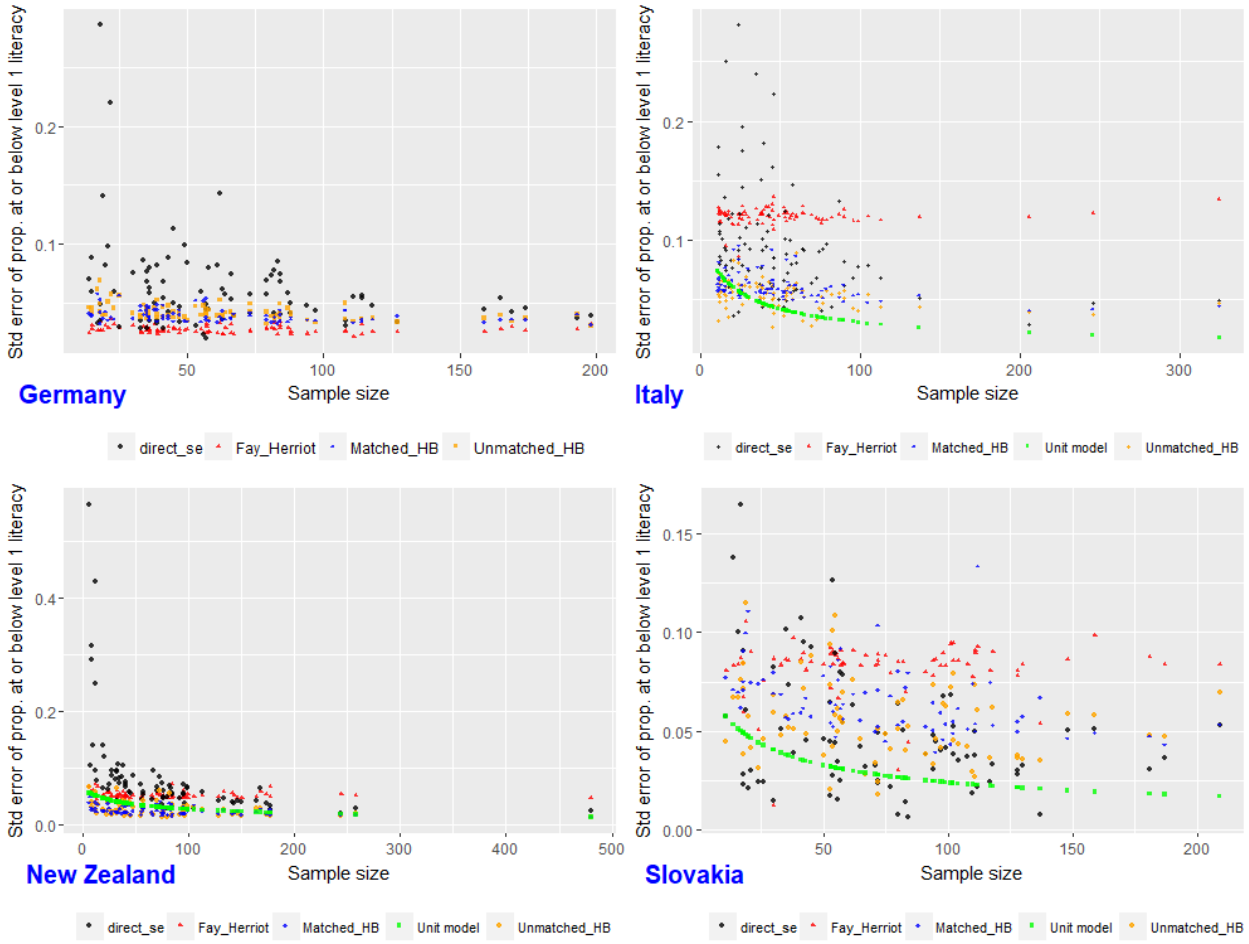


Figure 5-6 shows the standard errors of the direct estimates and the MSEs of all model estimates in one plot for each country. For these plots, keep in mind that the MSEs depend on the size of the estimated proportion. Therefore if the model proportion is different from the direct proportion, the variance will in theory be different, therefore the resulting MSE is not necessarily an improvement to the estimates due to the model. The MSE plot shows that almost all models produce smaller MSEs than the direct estimates, especially for areas of very small sample sizes. For Slovakia there appears to be less positive impact on the precision from the models. However, some investigation revealed that several direct estimates are close to zero and that SAEs have slightly higher values (likely due to shrinkage). As seen in the formula for the standard error of a proportion, the standard

errors for proportions are associated with the magnitude of the proportion, and therefore it is hard to make a conclusion as to the impact on standard errors, especially in the case of Slovakia's proportions.

In general, the unit-level EBLUP does not account for sample weight and design features. Therefore, the MSEs generated from the unit-level EBLUP models show strong correlation with sample sizes.

Figure 5-6 Proportion at or below Level 1: Comparison of standard errors between direct and SAE approaches



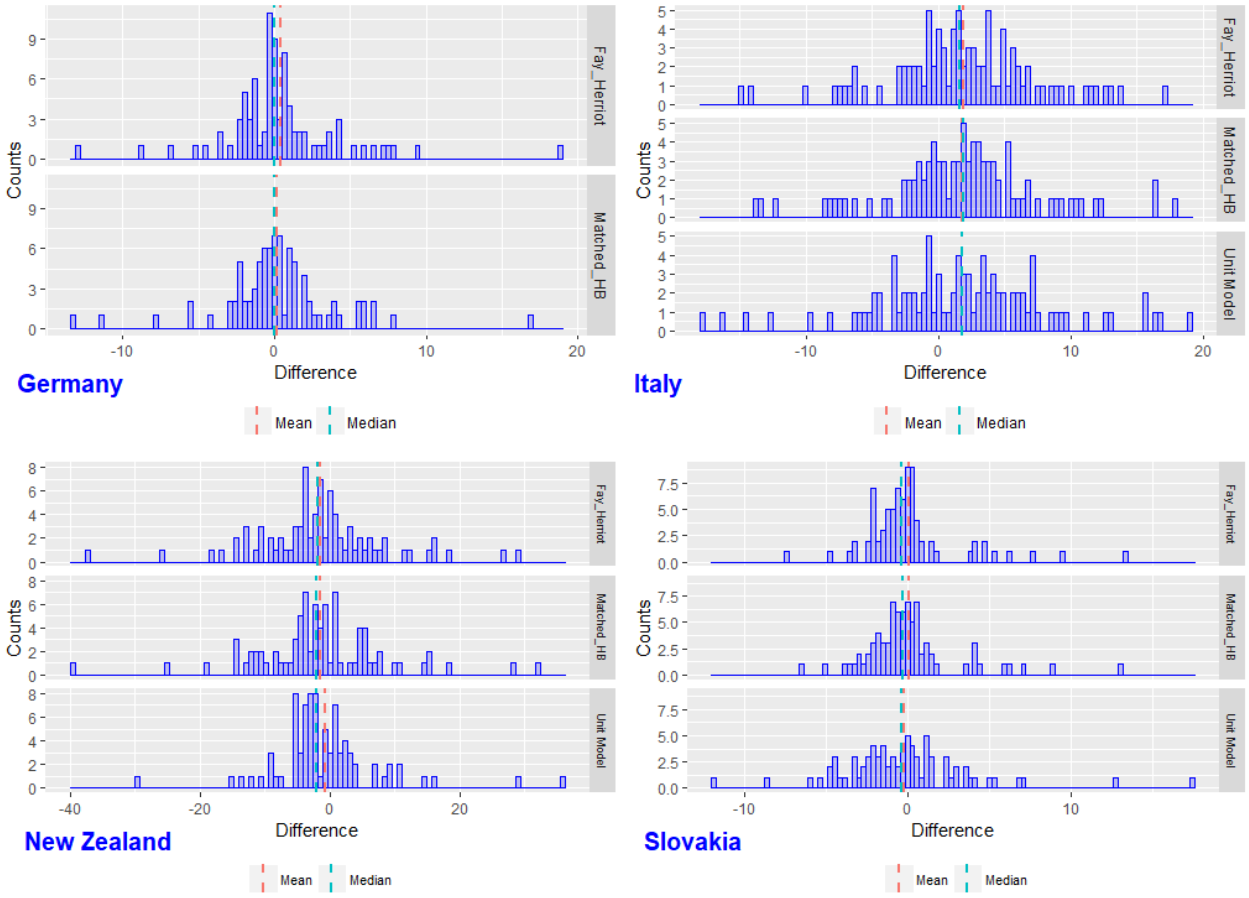
Averages

The same set of plots, as above, were generated for the average scores of literacy. The direct estimates of average scores have much smaller standard errors compared to the direct estimates of

proportions. As a result, the direct estimates of average scores are more reliable and make more contribution to the model estimates.

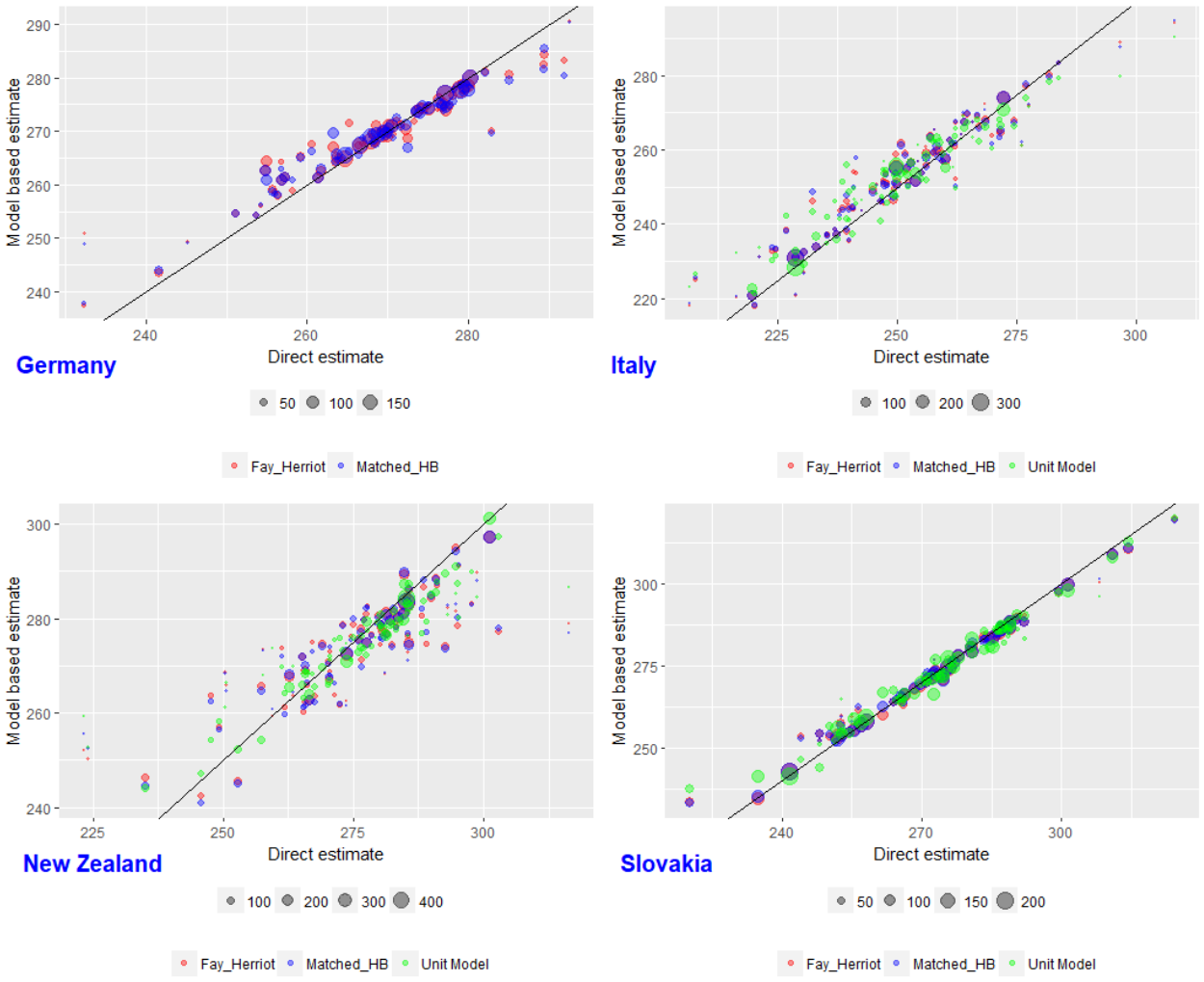
Figure 5-7 shows that the means and medians of differences between the model estimates and the direct estimates are almost zero for all four countries. The performance of the models of average scores look much better for Germany, than it was for the proportion at or below Level 1.

Figure 5-7 Average: Histogram of differences between SAE and direct estimate



In Figure 5-8, the bubbles are closely around the 45 degree lines for majority of the areas in four countries except for a few areas with small sample sizes. For Slovakia, the direct estimates and the model estimates are very similar.

Figure 5-8 Average: Scatterplot of SAE and direct estimates, with sample size as bubbles



Shrinkage towards the means can also be observed in Figure 5-9, but mainly for areas with very small sample sizes. Very little shrinkage is observed for Slovakia.

Figure 5-9 Average: Shrinkage plots of point estimates by sample size

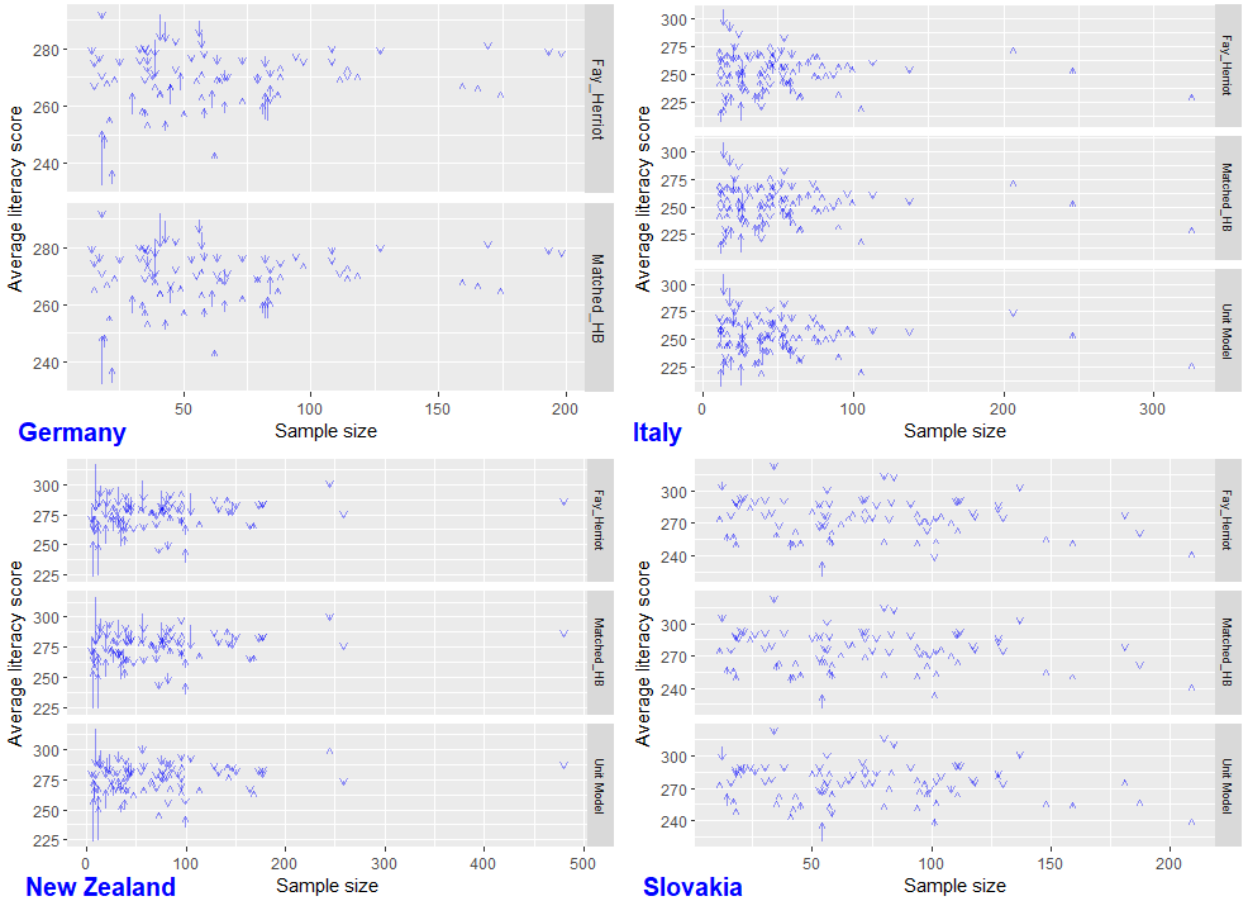


Figure 5-10 shows that the confidence/credible intervals of the SAEs from the various models almost always cover the direct estimates.

Figure 5-10 Average: Indication of coverage by confidence/credible interval

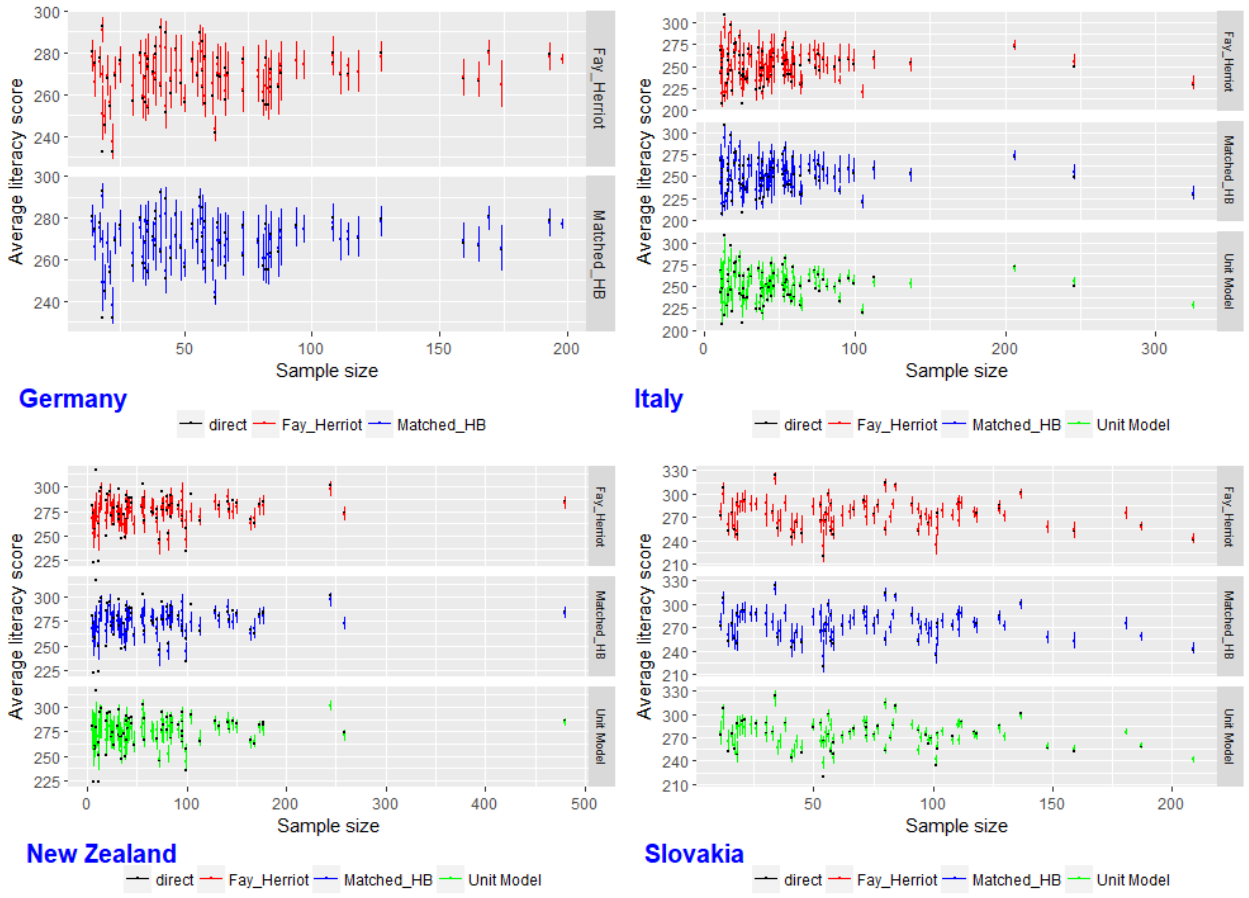


Figure 5-11 confirms that the model estimates are similar in most of the areas.

Figure 5-11 Average: Comparison of point estimates between direct and SAE approaches

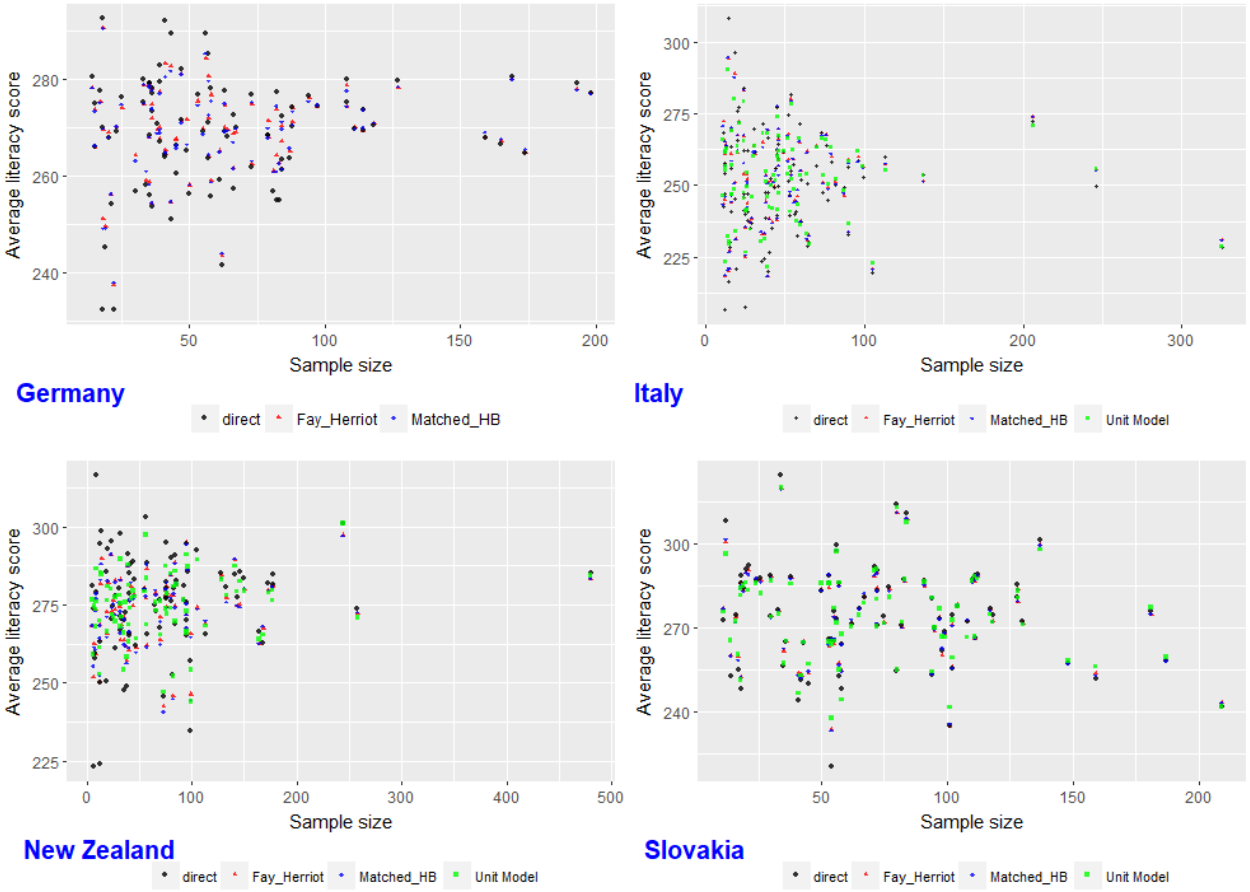
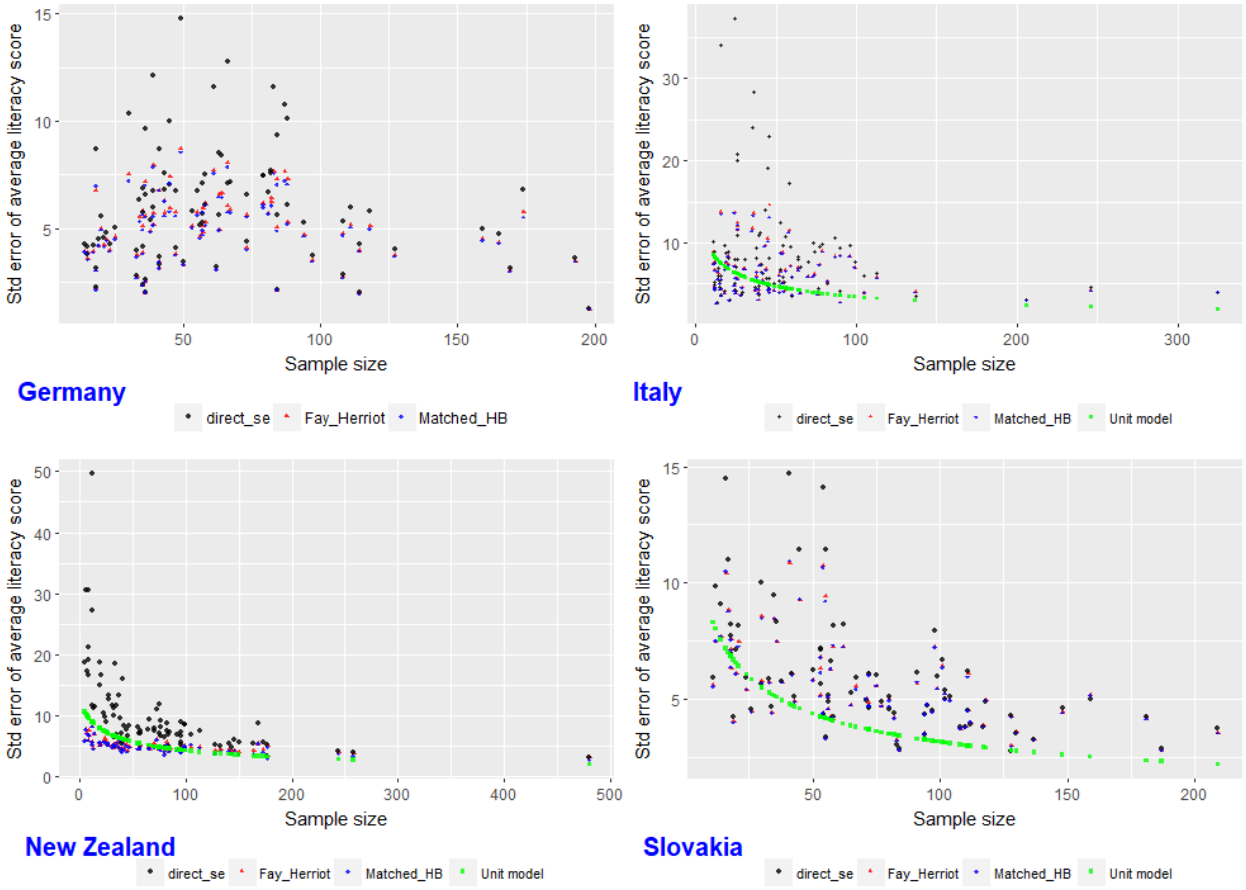


Figure 5-12 shows the model estimates are associated with smaller MSEs than the direct estimates, especially when the sample size is small in an area.

Figure 5-12 Average: Comparison of standard errors between direct and SAE approaches



5.3 Diagnostics and Evaluation Toward Publishable Small Area Estimates

The tools described in the prior section are only a small subset of diagnostic tools required to meet a standard level of quality for publishable SAEs. The above tools are aligned with the purposes of the initial research, however more thorough review is needed toward publishable estimates. This section includes references to some approaches that can be used to help ensure adequate quality in the final estimates.

5.3.1 Model Diagnostics

For model diagnostics to be carried out in Phase 2, we recommend internal validation and model sensitivity approaches. For internal model validation, examples of checks include the following.

- Checks for linearity of the relationship between predictors and target variables
- Checks of the distribution of residuals
- Checks of homogeneity of the variance (when checked against the predictor variables)
- Cross-validation
- Posterior predicted p-values (discussed in Rao and Molina (2015))
- A simple R^2 value of a multiple regression in the similar form of the linking model

The above validation consists of checking the model for its accuracy and robustness. A failure of these checks indicates that a revision of the model might be necessary (for example, adding new predictors, or nonlinear functions of the old predictors such as quadratic and interaction terms).

Examples of model diagnostics employed in SAE projects include Bijlsma, et al. (2017) checks on normality assumptions through the use of Q-Q plots, and Bauder, Luery, and Szelepka (2016) checks on standardized residuals. In Bauder et al. (2016), the averages of standardized squared residuals over large groups of observations were checked if they were close to zero, and checked for extremely small or large values.

In terms of model sensitivity, alternative models can be fit to the data to determine whether the model results were sensitive either to the prior distributions used for modeling or to the set of covariates used in the model. Such an analysis can support the choice of the final model showing the SAEs were not sensitive to the variants of the model that were investigated. The deviance information criterion (DIC) measure (Spiegelhalter, Best, Carlin, and Van der Linde, 2002) can be used to compare models.

5.3.2 Model Evaluation

The plots described in Section 5.1 are a selected sample of model evaluation tools. In addition, the model evaluation should include comparisons of aggregated SAEs to the country level.

Table 5-1 shows the average score and proportion at or below Level 1 for literacy estimated through three approaches: aggregated weighted SAEs (weighted by population total in each SA), national direct estimates using the International Data Explorer (IDE) and national direct estimates using the public use file (PUF). Both the SAE and PUF sources have imputed scores for literacy-related nonresponse (LRNR) cases by setting them equal to the first percentile, while the IDE had the score missing for LRNR cases. The table shows that SAE estimates are closer to PUF estimates than IDE estimates for all countries except Germany. In general, the estimates are similar for the three approaches.

Country	average literacy score			proportion ≤ literacy level 1		
	SAE*	IDE**	PUF*	SAE*	IDE**	PUF*
Germany	271	270	268	0.17	0.17	0.19
Italy	252	250	250	0.27	0.28	0.28
New Zealand	276	281	278	0.14	0.12	0.14
Slovak Republic	273	274	274	0.12	0.12	0.12
Sweden	278	279	279	0.13	0.14***	0.13

*The estimates for SAE and PUF include LRNR cases by assigning the first percentile of scores to LRNR
**The estimates from IDE do not include LRNR cases. Also the IDE estimates are the sum of the proportion less than Level 1 and the proportion at Level 1.
***Sweden does not have LRNR cases, so the IDE estimates should be the same as PUF estimates. The difference here is mainly caused by rounding since the IDE proportion ≤ Level 1 is the result of adding the proportion <Level 1 and proportion=Level 1.

The above analysis can be extended to compare to PIAAC estimates for regions and country-level estimates based on other characteristics derived as coarsened percentages of characteristics (e.g., estimated proportion in Level 1 or below in small areas in the lower third of percentage of the population with tertiary education). An example is given in Table 5-3 in Mohadjer, et al (2009). If there are differences between aggregated indirect estimates and country-level direct estimates, benchmarking could be done to adjust the SAEs.

6. Summary of Phase I Research Results and Recommendations for Phase 2 (Production of SAEs)

The goal of the initial research was to evaluate the feasibility of producing local area estimates using PIAAC data. To accomplish this goal, SAE models were fit to PIAAC data from a group of countries with various sizes and different sample designs. The following provides our general findings about the process and the modeling aspects, and includes the specific conclusions for each country participating in this initial phase.

6.1 General Findings

The following highlights the general findings of the initial SAE research phase:

Input data process can be improved. The input data process was time consuming, and prone to error. The process can be improved by emphasizing and formalizing the standard input data file layout and guidance. There were a number of other lessons learned that would need to be addressed. For example, external source variables that did not match in definition to the PIAAC survey variables could not be used in the SRE process or the unit-level SAE modeling. Area-level models provide a way to use such information, even though the definitions do not match. That being said, the initial research results can be improved upon by having more variables in SRE and/or unit-level SAE models. The following are our initial general guidelines for countries to follow when submitting the input files to Westat :

- Provide a mapping of the PIAAC covariates to the population covariates;
- Collapse covariates so that they have a maximum of three or four categories;
- Use consistent variable names and file layouts; and
- Review definitions and distributions of covariates in the PIAAC data file and population file for consistency, and possibly create recoded variables in their PIAAC file that match the census categories.

Covariate identification process can be improved. Westat provided some guidance as to what topics (e.g., age, education) to cover by the covariates, and countries were in charge of identifying

the source of the covariates in their countries. The covariate identification process ended up to be quite limited for most countries. Given the importance of covariate data in modeling, Westat guidance can be improved to include recommendations on where to look for appropriate covariates and in further assisting countries on the choice of the covariates. Correlations, similar to what is shown in Tables 3-2 and 4-1, could be checked and reviewed by countries prior to deciding which covariates to submit to Westat.

In the initial research, the covariate pool was limited, and the predictor search was deterministic (in lieu of analytical) due to the limited choices available. We should note that if the covariate estimates are noisy, then alternative sources should be used (such as registry data or census), or the work from Ybarra and Lohr (2008) should be incorporated to address the measurement error from the covariate estimates. In general, the covariates used in this research tend to have good association with the direct estimates. In addition, because of the estimation of the coefficients in the SAE models, it is good to limit the number of predictor variables in a SAE model. For example, if there are only 30 SAs, then having more than 3 predictors in the model could cause significant instability in the model predictions. Therefore, because of the reliance on good covariate information, we think it is reasonable to only conduct SAE with 30 or more SAs. A general rule of thumb is that there needs to be 10 to 30 data points for each term of the SAE model. For example, suppose there are 100 SAs, the SAE producer would only identify 3-to-10 predictors for the SAE model. Using this rule of thumb, there may be slightly fewer predictors used in the Phase 2 models than what was used in Phase 1.

Area-level models have good potential for providing reliable PIAAC SAEs for all countries. This is evident by the results, showing that the SAEs are close to direct estimates for SAs with the largest sample sizes, and the SAE models showed impact on SAs with the smallest sample sizes. Also, in general, the MSE plots show that almost all the models produce smaller MSEs than the direct estimates, especially for areas with very small sample sizes. From the evaluation plots, the unmatched and matched HB models have the most potential, and also did well for averages. Area-level models can rely on area-level data, which may be the only data available. The unit-level model has potential for better estimates for countries with a wealth of registry data and without clustering within their SAEs. One can see the direct association of the standard errors to the sample size for unit-level models in Figure 5-6 for example, illustrating the effect of ignoring the design effect.

Further research and development is needed on performing diagnostics and evaluating model results. We observed that some improvements could be made to the resulting model fit. Such results demand more investigation as to why the fit of the model is in question, and to improve the model fit through transformations, combining categories of variables in different ways, and introducing interaction terms. Graphical checks were done on the model predictions for the initial research, however, more thorough checks are needed to validate the results. Several recommendations are provided in Section 5.3.

The SAE process is feasible for PIAAC. To a certain extent, software was developed that ties together the various sub-processes. More extensive data checks are needed, and more generalization of the programs is needed to facilitate model development for all countries. SAS Proc MCMC was used to generate SAEs for the matched and unmatched HB area-level models, and SAS Proc MIXED was used for the unit-level model in combination with Proc IML to compute the MSEs. Assisted by these computer procedures, the research showed that the SAE process is certainly feasible to do, however, some challenges need to be addressed as noted in this section. Extensions to the models to include another random effect, or multivariate models, are also feasible.

Country feedback was positive in general. See discussion given in Section 6.2.

Potential for an improved precision measure through parallel PV processing. While the initial research had in it many of the steps needed to capture the various key sources of error, further consideration of parallel PV processing can improve the handling of imputation error variance.

Guidance for model choice has been developed. One outcome from the initial research was to identify scenarios for each country that factor into the decision about the SAE model framework for Phase 2 (creating publishable SAEs). These factors are 1) whether external covariate information is available for which the variables match the survey item definitions, 2) whether design contains informative sampling both in terms of clustering within the SAs and variation in the weights, and 3) if the estimated proportion is on average less than 0.20 or not. Table 6-1 provides the various scenarios and a recommended model to use for SAE. As seen in the initial research, the recommended model is not necessarily as clear-cut as Table 6-1 shows, and therefore, some investigation of the model-type choice is typically needed in practice.

Table 6-1 SAE scenarios and recommended model type

Scenario	Covariates match survey item definitions?	Informative sampling		Estimated proportion < 0.2?	Recommend	
		Clustering within small areas?	Differential weights?		Proportions	Averages
1	Y	Y	Y	Y	ALU	ALM
2	Y	Y	Y	N	ALM	ALM
3	Y	Y	N	Y	ALU	ALM
4	Y	Y	N	N	ALM	ALM
5	Y	N	Y	Y	PE*	ALM
6	Y	N	Y	N	PE*	ALM
7	Y	N	N	Y	UL*	UL
8	Y	N	N	N	UL*	UL
9	N	Y	Y	Y	ALU	ALM
10	N	Y	Y	N	ALM	ALM
11	N	Y	N	Y	ALU	ALM
12	N	Y	N	N	ALM	ALM
13	N	N	Y	Y	ALU	ALM
14	N	N	Y	N	ALM	ALM
15	N	N	N	Y	ALU	ALM
16	N	N	N	N	ALM	ALM

*Note: If non-linear model, then a full cross-tab of covariates is needed at small area level.

ALU = Area-level unmatched model

ALM = Area-level matched model

UL = Unit-level model

PE = Pseudo-EBLUP

6.2 Country-specific Observations and Recommendations

As part of the research process, SAEs of proportions at or below Level 1 on the literacy scale, and average literacy score were provided to countries, along with summaries of the process as given in Appendix C[9]. The research, including general feedback from countries, showed that the predictions can potentially improve upon the direct estimates with small sample sizes by borrowing strength from the covariates. Countries provided insights and questioned when a particular estimate was higher or lower than the direct estimate. In those cases, Westat worked with the countries to review the sample size, the confidence intervals, and to ensure the precision of the estimates were taken into account. The covariates can also be reviewed to see the data that went into the models. A good

[9] Small updates were made after countries reviewed, for example, Sweden's information was updated to reflect that a unit-level EBLUP was applied.

example is given below for Italy. The following paragraph provides some highlights of the results of the initial research for each country.

Germany. Although the input data process and modeling process went smoothly, there is some evidence of lack of model fit from the evaluation. The results look better for averages than for the proportion at or below Level 1. For proportion at or below Level 1, the bubble plot in Figure 5-2 shows that the model estimates are usually smaller than the direct estimates when the estimated proportions are larger than 20 percent. There is also some indication of an influential observation, as noted in Section 5.2, and possible lack of model fit as illustrated by the Fay-Herriot model results being smaller than the direct estimates on average.

We think the main challenge is to improve the covariate selection process by focusing on bringing in stronger predictor variables for the model. The current set of covariates have weak pairwise correlations between the direct estimates and covariates. In general, in the absence of good covariates, SAE modeling for proportion at or below Level 1 may not be possible, and more sample would be needed to achieve adequately representative model-assisted (via SRE) direct estimates. However, more work could be done to improve the models by combining categories of variables, looking for the need of interaction terms, or transformations.

Germany raised that the precision of the covariate estimates should be reviewed, because they are generated from the micro-census, which points to the presence of potential non-negligible associated sampling error. If the covariate estimates are too noisy to be considered as model predictors, then other data sources should be used (such as registry data or census), or, as mentioned above, the work from Ybarra and Lohr (2008) should be incorporated to address the measurement error from the covariate estimates. The review of the precision levels showed negligible sampling error in the covariates.

Another challenge, imposed by confidentiality and data sharing restrictions, limited the modeling options that could be applied to the Germany data. For example, the SRE approach could not be applied due to unavailability of unit-level covariates because of confidentiality reasons. Refer to Section 7 for a more general discussion on data sharing.

To proceed to the second phase, Germany would need to locate and submit stronger predictor variables, and search for interaction terms and transformations, for the SAE model. In addition, they

need to conduct the SRE step prior to submitting the direct estimates to Westat (see Section 2 describing the process of data submission by Germany).

Given the informative PIAAC sample design and weighting process (addressing nonresponse bias, other coverage error, and the clustering), Westat would consider an area-level matched or unmatched model for Germany. Germany would also like to produce SAEs for their Federal states. This can be accomplished with a similar two-fold random effect model as employed during the NAAL SAE process.

Italy. The process of data submission and modeling went smoothly, except for the required additional steps conducted by Westat for creating the final input files. The submitted input data included covariates in the population file with many categories, requiring additional work in creating combined categories. Future data file requests should include guidelines on the limits on number of categories for each covariate.

The pairwise correlations show a number of strong covariates, and the results of SAE evaluation looked good in general. Italy pointed out a region with mixed results for the proportion at or below Level 1: SRE = 20%, Direct = 21%, F-H area level = 18%, Unmatched HB area level = 11%, Matched HB area level = 16%, Unit level = 22%. There are several points to make about these results. The SRE and direct estimates are only based on 11 cases for the region, resulting in unstable estimates. For the example region, the three area-level models are all producing lower estimates. The covariates are fairly strong and there are two impactful variables (citizenship and marital status) used in the area models that were not used in the unit-level model (which has a higher estimate). The linear SAE area models (F-H and Matched HB) are similar. The unmatched HB area model tends to handle proportions less than 0.2 better than the linear models.

Given the informative PIAAC sample design and weighting process (addressing nonresponse bias, and other coverage errors), Westat would consider using an area-level matched or unmatched model for Italy. Italy would like to produce SAEs for their NUTS2 regions. This can be accomplished with a similar two-fold random effect model as employed during the NAAL SAE process.

New Zealand. The SAE process went smoothly for New Zealand's data in general, and the covariates as a whole had good association with the PIAAC direct estimates. One challenge could be confidentiality and data sharing of input variables. For example, some missing data existed in the

input data file of covariates because of confidentiality restrictions for sharing data on rare combinations of covariates. Therefore an imputation process was employed to derive imputed values prior to use in the SRE modeling. The preparation of the input files and the imputation process was very time consuming. Our general guideline for Phase 2 would be for New Zealand to provide a full cross-tabulation of the covariates.

The following is a summary of the observations made by New Zealand upon review of the SAEs for the proportion at or below Level 1.

- For proportions at or below Level 1, the SREs and the direct estimates are similar to each other for all the small areas – as expected.
- For several small areas, the direct estimates of the proportion at level 1 literacy or below look implausibly small or large – based on the co-variates and also knowledge of the socio-demographic character of the areas. This compares with the more believable values from the four models.
- For the small areas with quite large PIAAC samples, the models provide values similar to the direct estimates and similar to each other.
- A couple of areas were scrutinized further, and determined that based on the model values seeming more believable than the direct estimate for a range of areas, these model values would be accepted too.
- Across the small areas, the sets of values from the four models are diverse in pattern, including all four are quite close, to all four quite different.

New Zealand concluded that in terms of the proportion at or below Level 1, the models' use of Census data seems to generate better estimates of the proportions of people with low literacy than the direct estimates or SRE using just the survey data, and much better estimates for areas with small PIAAC samples.

Given the informative PIAAC sample design and weighting process (addressing nonresponse bias, other coverage error, and the clustering), Westat would consider an area-level matched or unmatched model for New Zealand.

Slovakia. The modeling process went well in general, and the covariates appeared strong. However, we had similar challenges with the input data files as we had with Italy and New Zealand. It took

time to re-arrange the population files into one table. The covariates had a lot of levels (smallest was 8 levels) that required collapsing prior to any modeling stage.

Slovakia indicated that the proportion of the population at or below Level 1 in literacy is higher in Presov region (small area ID starting with number 7) according to their regional analysis and that seemed to be in-line with the SAE estimates. We note that the impact on the standard errors for the proportion at or below Level 1 is hard to determine due to some very small direct estimates and the variance directly associated with the estimated proportion (whether direct or modeled), as discussed in Section 5.2. In general, the improvement from the models is more evident in the standard errors for averages. We note also that for the averages, the plot in Figure 3-4 shows that most areas had SRE estimates lower than the direct estimates.

Given the informative PIAAC sample design and weighting process (addressing nonresponse bias, other coverage error, and the clustering), Westat would consider using an area-level matched or unmatched model for Slovakia.

Sweden. SAE modeling was limited to an attempt at using a unit-level EBLUP model, which was able to be applied to Sweden data even though the number of SAs was small (8 SAs). As with other countries, the SRE was also applied as a model-assisted direct estimation approach. In terms of providing input data, there was some difficulty with the ISCED equivalence. For Phase 2, countries will be asked to provide a detailed description of what each covariate category means, and how it relates to PIAAC variable categories.

For the proportion at or below Level 1, in general, the SREs and direct estimates were similar. Sweden expected that the SREs would be higher than the direct estimates because individuals with low education are underrepresented among respondents. There were two regions where differences were observed, one where the direct estimate was lower and one where the direct estimate was higher. These are the type of investigations that are needed prior to publishing SAEs to understand the reasons for the results, and to make sure the results make sense. For the averages, Figure 3-2 shows that seven of the eight areas had SRE estimates lower than the direct estimates.

If Sweden is interested in producing SAEs for more SAs, and is able to share input data at the SA level (need PIAAC respondents with SA identifier), then Westat would consider either a unit-level EBLUP model, or a pseudo-EBLUP (You and Rao, 2002; Rao and Molina, 2015), given their

unclustered design (acknowledging the informative weights). Otherwise, Westat can further investigate the unit-level model for the 8 SAs, or would consider an area-level SAE model for the 21 SAs. For an area-level model, the input data would need, at a minimum, 1) the PIAAC direct estimates and standard errors for each small area, and 2) for each covariate, the univariate percentages for key characteristics for each small area (e.g., percentage unemployed). Therefore no microdata would be needed under this scenario. In either case of area- or unit-level models, Westat could consider a two-fold random effect model as employed during the NAAL SAE process, with the set of 8 regions and the set of 21 regions represented by the two random effects.

7. Review of Critical Components for Countries Interested in SAE

Sections 2 to 6 described the research and evaluation of various SAE approaches applied to five countries, and pointed to the experiences gained in this process. This section focuses on the critical points that impacted the SAE research, and thus needs to be addressed before any work can proceed on the production of small area estimates for PIAAC countries. The main issues are around the quality and accessibility of the PIAAC data and the covariates at the local area level.

The following highlights some of the critical aspects that countries interested in SAE would need to consider.

Statistic(s) of Interest. For this research activity, we focused on the proportion of the target population at or below Level 1 of literacy, and the average score for literacy. Countries may choose other statistics, such as proportion of the target population at Level 2, or at Level 3 or above, or consider numeracy statistics, for example. Countries can consider multivariate estimation approaches in which several population distributions can be estimated using a single multivariate model.

Local Areas (Geographic Units) of Interest. One of the first steps in SAE is to determine the local areas for which small area estimates are desired. Data users naturally want data for as small a geographic unit as possible. OECD has expressed interest in looking at NUTS Level 2 and NUTS Level 3. This is a critical first step as it has implications on the SAE development as well as potentially impacting the national PIAAC sample design (as described below). Therefore, countries need to consider a number of factors when defining their local areas of interest. These factors are described below. Finally, countries can consider multiple levels of small areas, as long as the smaller areas are contained within the larger ones (e.g., counties, states, regions). This can be incorporated in SAE through the introduction of higher-level random effects.

Suitability of the PIAAC Sample for SAE. Whether the PIAAC national sample is suitable for SAE depends on the definition of the local areas, the distribution of the population across the local areas, and the diversity of the demographic characteristics of the local areas. The PIAAC technical standards and guidelines (TSGs) goals are to produce the most efficient samples for the production of the national estimates, given the costs of data collection. That is, the TSGs require that countries

design and select the PIAAC samples with the goal of arriving at the most optimum sample (minimum mean square error) for the production of national estimates, given the costs of data collection and the specific sampling conditions in each country. A sample design that is optimum for national estimates is rarely also optimum for estimation at the small area level since national samples are representative of the population, whereas a sample that represents the small areas will be the most efficient for SAE. Thus, a critical first step is to ensure that the national PIAAC samples have adequate representation of the small areas of interest.

Westat conducted an evaluation of the suitability of the PIAAC sample for SAE for the five countries involved in the research. In addition, we conducted a similar evaluation for the US PIAAC prior to start of the Round 3 data collection in the US. The method used to evaluate the coverage of the sample with respect to the small areas was to compare the sampled small areas with their population distribution by the demographic characteristics that are highly correlated with proficiency.

All five countries involved in the Phase 1 research had a good coverage of all types of small areas in their PIAAC sample. This is partly evident by the distributions provided in Table 2-3, showing that all countries had PIAAC samples in most of their small areas. In addition, further analyses, as shown in Table 7-1, confirmed that small areas with various levels of proficiency (as indicative by their demographic characteristics) had good representation in the PIAAC samples. This may not be the case for other countries. For example, a similar analysis of small areas in the US PIAAC showed that the Round 1 and 2 samples (combined) were not adequately representing different types of small areas, and thus a third round of data collection was necessary to supplement the first two rounds to allow SAE to proceed.

Table 7-1 Coverage of PIAAC small areas by area characteristics

Country	Coverage of Small Areas
Germany	Three variables (education attainment, employment status and nationality) were each split into two levels (high, low) based on their distribution of SAs in the population. A 3-way cross-tabulation of SAs (2x2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For Germany, the coverage is 8 out of 8 cells even after excluding SAs with less than 30 cases.
Italy	Three variables (education attainment, employment status and citizenship) were each split into two levels (high, low) based on their distribution of SAs in the population. A 3-way cross-tabulation of SAs (2x2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For Italy, the coverage is 8 out of 8 cells. When excluding SAs with less than 30 cases, the coverage is 7 out of 8 cells. Improvement to the SAEs can be made by adjusting the sample design to cover all cells with sample sizes larger than 30.
New Zealand	Three variables (education attainment, employment status and birthplace) were each split into two levels (high, low) based on their distribution of SAs in the population. A 3-way cross-tabulation of SAs (2x2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For New Zealand, the coverage is 8 out of 8 cells. When excluding SAs with less than 30 cases, the coverage is still 8 out of 8 cells. Improvement to the SAEs can be made by adjusting the sample design to cover cells of a 4-way table, by adding another key variable to the cross-tabulation.
Slovakia	Three variables (education attainment, employment status and nationality) were each split into two levels (high, low) based on their distribution of SAs in the population. A 3-way cross-tabulation of SAs (2x2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For Slovakia, the coverage is 8 out of 8 cells. When excluding SAs with less than 30 cases, the coverage is 7 out of 8 cells. Improvement to the SAEs can be made by adjusting the sample design to cover all cells with sample sizes larger than 30.
Sweden	Two variables (education attainment, and birthplace) were each split into two levels (high, low) based on their distribution of SAs in the population. A 2-way cross-tabulation of SAs (2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For Sweden, the coverage is 4 out of 4 cells. When excluding SAs with less than 30 cases, the coverage is still 4 out of 4 cells. If more SAEs are considered in the future, improvement to the SAEs can be made by adjusting the sample design to cover cells of a 3- or 4-way table, by adding one or more key variables to the cross-tabulation.

We will follow a similar evaluation approach for other PIAAC countries interested in SAE. That is, we will review the distribution of the sampled small areas as compared to the population of the small areas with respect to the characteristics correlated with proficiency and examine whether different types of small areas are well represented in the PIAAC sample. If the evaluation shows that the national sample does not have a satisfactory representation of the small areas, then country has two options; 1) revise the definition of small areas, or 2) consider increasing the sample size for PIAAC (beyond the minimum required sample size) to improve coverage of the small areas. The

objective of option 2 is to arrive at more accurate direct estimates for local areas with different proficiency levels (as reflected through the demographics of the local areas).

We will be open to discuss various approaches in increasing the sample size (option 2) that could serve other purposes, in addition to SAE. For example, US decided to select a nationally representative sample for Round 3, and at the same time minimized overlap with the Round 1 and 2 samples. Following this approach gave them the opportunity to create mid-decade (2017) national estimates for PIAAC as well as producing small area estimates by combining Rounds 1, 2 and 3 from the first cycle of PIAAC.

Small Area Sample Size. Another aspect of the PIAAC national sample that impacts SAE modeling is the number of small areas with PIAAC sample. As mentioned in Section 2, four out of five countries that participated in this research had over 30 (used as the minimum sample size in this research) small areas with adequate within SA sample size for the derivation of the direct estimates. Facing data-sharing restriction challenges, Sweden decided to reduce their number of small areas to eight areas. The small sample size prohibited any type of area-level SAE model fitting. Therefore, a model-assisted direct estimation approach was used for Sweden, as well as a unit-level EBLUP model. In general, we expect the following rule to be applicable across countries:

- PIAAC includes more than 30 small areas with samples greater than 30 completed assessments.
  - SAE modeling can proceed given all other requirements (as mentioned below) are satisfied.
- PIAAC includes between 20 and 30 small areas with samples greater than 30 completed assessments.
  - Further evaluation is required for such a small sample size.
  - It will be dependent on the quality of covariates and the correlation with outcome variables.
- PIAAC includes samples in less than 20 small areas.
  - Model-assisted direct estimation will be applied to areas with at least 100 sample cases.
  - Unit-level modeling might be feasible in certain situations.

Covariates. A critical aspect of SAE is the availability of covariate data at the small-area level so that appropriate SAE models can be developed. In addition, countries need to verify that the covariate data satisfy the following conditions to ensure the SAE model produces reliable estimates.

- It is ideal that covariates have an exact matching of definitions with those in PIAAC (i.e., the exact matching of the definitions is required for the SRE approach and for the SAE unit-level modeling but not a necessary requirement for the SAE area-level modeling).
- They are highly reliable, i.e., coming from censuses or very large surveys such as the labor force surveys.
- The data is up-to-date, ideally matching the PIAAC data collection dates.
- They have proven to be highly correlated with adult proficiency.

An important step is for countries to conduct a thorough search of all available covariates that are hypothetically associated with literacy. Countries should search their census, registries, and largest surveys for local area estimates for covariates. There could potentially be dozens, if not over 100 covariates from various data sources, but the eligible covariates for the predictor selection process should have very small measurement error (or sampling variances) associated with them. Westat would then compute survey regression estimates for each SA, and then establish the final predictors for the model through a predictor search process. (An analytical-driven predictor search process was not conducted in the initial research.)

Sharing Data with Institutions Outside the Country. As mentioned is Section 2.1, a limitation on the research was imposed by the country's ability to share data with Westat, whether the data was internal or external to PIAAC. With the matched HB model performing well in the research, the covariate data may only need to be area-level univariate counts (not crossed with other covariates). Therefore, confidentiality may not be a road-block for most countries where the matched HB model is the most appropriate model. However, limited access to covariate data due to confidentiality can still have some implications, as mentioned specifically for Germany and New Zealand in Section 6.2. Although an area-level model can be applied, limitations due to confidentiality may impact the quality of the results, for example, the unit-level SRE step in the process may not be doable.

In certain cases, one critical initial step for countries is to find out whether they can share both the PIAAC data and the external covariate data at the local level with Westat or other institutions

outside their border. In the case where country is not allowed to share data at the defined local area level, then the next step is to decide whether the country is willing to redefine the local areas to an aggregated level. Our experience showed that countries were allowed to share data when they redefined their small areas to a level at which data sharing was allowed, and that the aggregated SA was still of interest to the country.

Sorting out issues of confidentiality may be addressed on a country-by-country basis. For example, if confidentiality restrictions prohibit a country to share PIAAC data or any covariates at the SA level, then the most appropriate option might be for the country to run the steps that involve the confidential data in the country (in light of national standards and the General Data Protection Regulation (GDPR)), and submit the anonymized outcome data files to Westat for the remaining steps of the process including diagnostics and model evaluations. We should point out that the steps involved in preparation for model development can be quite complex and will require an expert in the area of statistical modeling within the country. Consulting with Westat can provide insights on the impact of IRT modeling and estimation of PVs, informative PIAAC sampling and nonresponse (including proficiency-related nonresponse), as well as experience in working with other countries in PIAAC SAE estimation. Efficient processing via developing standardized software may also prove beneficial.

Figure 7-1 shows the overall process for the SAE activities, as highlighted by the country and Westat Tasks.

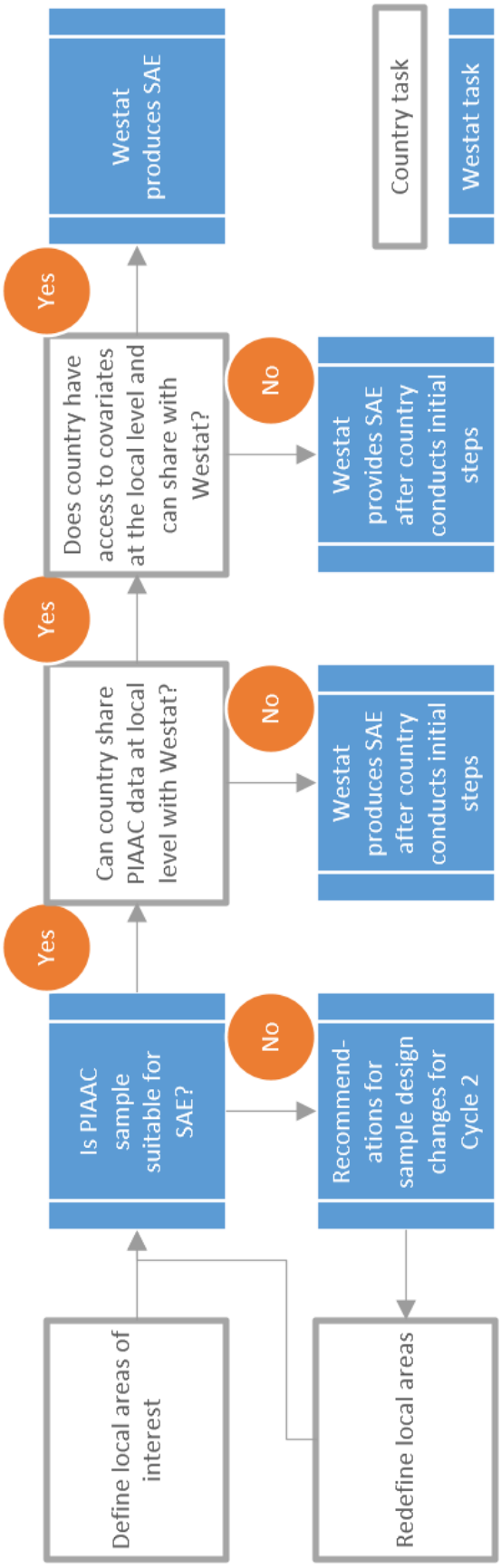


Figure 7-1 PIAAC SAE process chart

References

Bijlsma, I., van den Brakel, J., van der Velden, R., and Allen, J. (2017). Estimating literacy levels at a detailed regional level: An application using Dutch data. *ROA Research Memorandum ROA-RM-2017/6*. Maastricht, Netherlands: Research Centre for Education and the Labour Market.

Erciulescu, A., Cruze, N., and Nandram, B. (2017, August). Small area estimates for end-of-season agriculture quantities. [PowerPoint slides]. *Joint Statistical Meetings*, Survey Research Methods Section. Baltimore, MD: American Statistical Association.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*, 269–277.

Khan, D., Wei, R., He, Y., Shin, H-C, and Malec, D. (2018). Bayesian state-level estimates of diabetes prevalence in the United States 2006-2015. Presented at the Federal Commission of Statistical Methodology conference on March 7, 2018, Washington, D.C. Presentation slides available at http://www.copafs.org/UserFiles/file/2018FCSM/C-4Khan_March7_2018FIN.pdf

Lahiri, P., and Suntornchost, J. (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B: The Indian Journal of Statistics,* 77: 312. DOI 10.1007/s13571-015-0096-0.

Mohadjer, L., Kalton, G., Krenzke, T., Liu, B., Van de Kerckhove, W., Li, L., Sherman, D., Dillman, J., Rao, J.N.K., and White, S. (2009). *National assessment of adult literacy indirect county and state estimates of the percentage of adults at the lowest literacy level for 1992 and 2003* (NCES 2009-482). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T., and Van de Kerckhove, W. (2011). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics*, *66*(1), 1-9.

Organization for Economic Cooperation and Development (OECD). (2016). *Technical report of the survey of adult skills (PIAAC), 2nd Edition.* Paris, France: OECD. Retrieved from

https://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf

Pfefferman, D. (2013). New important developments in small area estimation. *Statistical Science, 28*, 40-68.

Rao, J.N.K. (2003). *Small area estimation* (Wiley Series in Survey Methodology). New York: Wiley.

Rao, J.N.K., and Molina, I. (2015). *Small area estimation.* Second Edition. (Wiley Series in Survey Methodology). Hoboken, New Jersey: Wiley.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B, 64*, 583-640.

Stukel, D., and Rao, J.N.K. (1999). Small-area estimation under two-fold nested errors regression models. *Journal of Statistical Planning and Inference*, *78*, 131-147.

Torabi, M. and Rao, J.N.K. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, *127*, 36–55.

Yamamoto, K. (2014). *Using PIAAC data to produce regional estimates.* [Unpublished manuscript]. Princeton, N.J.: Educational Testing Service.

Ybarra, L.M.R., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, *95*(4), 919-931.

You, Y. and Rao, J. N. K. (2002), A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30, 431–439.

Appendix A - Input Files A and B Data File Structure Requirements Submitted to Countries

The best predictors are covariates that are highly correlated with adult proficiency, such as gender, age, educational attainment, labour force status, and migration status. Countries needed to assess what population data items were available for submitting to Westat, as well as submitting PIAAC data at the local area level. Table 1 contains the instructions given to countries for providing input data needed (from the countries) to process the planned models.

Table 1 SAE input data from the country

Item #	Item	Description	Model type		Records	Note
			Unit	Area		
1	PERSID	PIAAC person ID	✓	✓	Respondents	
2	Wij	Sample weights	✓	✓	Respondents	
3	RWij	Replicate weights		✓	Respondents	
4	VCij	Variance Cluster identifier		✓	Respondents	
5	Yij	Literacy scores (10 plausible values)	✓	✓	Respondents	
6	Xij	Person level covariates	✓		Universe (all j)	Optional, for all j for prediction
7	Xi_bar	Small area level covariates (area level averages or proportions)	Optional	✓	Universe (all i)	For all i for prediction
8	AREAIDi	ID of small area level i (smallest e.g., counties)	✓	✓	Universe (all i)	Required for all i for prediction
9	AREAIDh	ID of small area level h (higher aggregate level – e.g., states)	✓	✓	Universe (all hi)	Optional, for all hi for prediction

h denotes high aggregate area

i denotes small area of interest

j denotes person

Notes:

- Countries needed to provide two files: FILEA = person-level file of respondents with Item #s 1 through 5, 8 and 9 (optional); and FILEB = covariates file that covers the universe of persons j, or areas i (discussed further in the following notes). FILEB needed to include 6, 7, 8, and 9 (where 6 and 9 are optional)
- For FILEA, the replicate weights (Item #3) and the variance clusters (Item #4) were used to compute the estimated variances for the area-level direct estimates.
- For FILEB, the person-level covariates (Item #4) suggested by Westat included

- Age
- Gender
- Race-ethnicity
- Education attainment
- Employment status
- Poverty status
- Born in or outside the country

Countries were welcome to provide other available covariates which are considered highly predictable for the estimates of interest. Examples of Xi_bar include percentages of the population for the small area i, such as: percentage of population age 60 and older, percentage of the population who are male, percentage of population who are Hispanic, percentage of the population with a college degree, percentage of the population who are unemployed, percentage of the population in poverty, and percentage of the population who do not speak the official language(s).

For FILEB, the person-level covariates (Item #6) were saved in a person-level file, or an aggregation file that takes the form of a cross-classification table with the population counts.

Small Area ID	X1	...	Xp	Population Count
1	1		1	100
1	1		2	50
...				
i	1		1	200
...				

- If some of the covariates were not available in the form of a person-level file, or cross-classification table, as an alternative, the countries were told to provide the corresponding area-level covariates (as means or proportions) listed in Item #7.
- The covariates in Item #6 and Item #7 could have been extracted or derived from administrative data or large-scale surveys.

Person-level covariates Xij in Item #6 were expected have the same definitions and categories as the corresponding PIAAC variables in order to fit a unit level area.

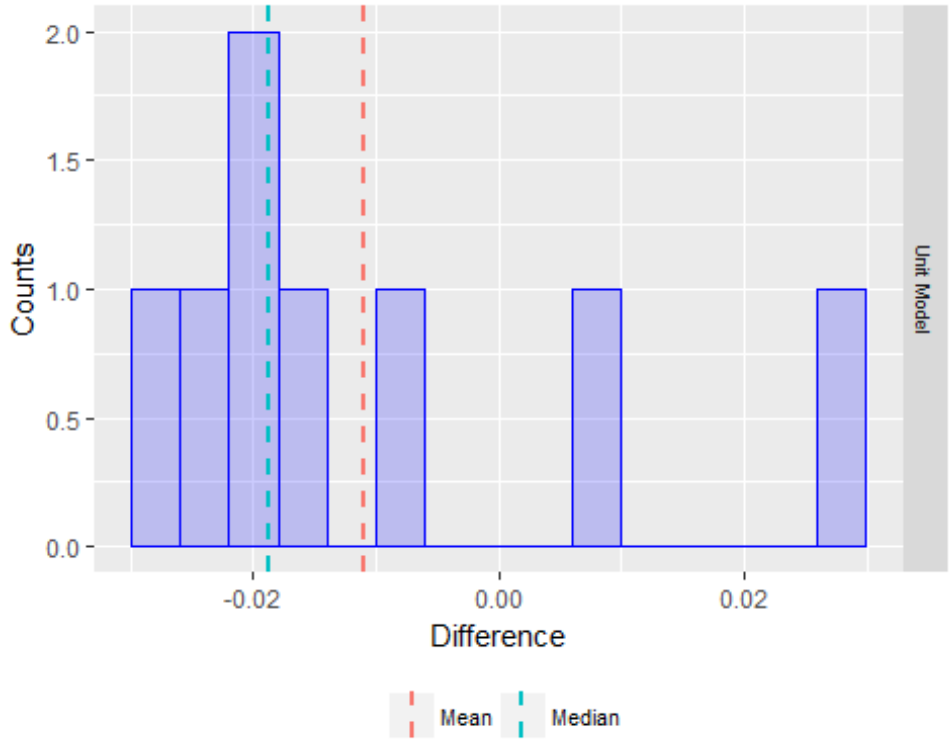
Appendix B - Evaluation Graphs for Sweden's Unit-level SAE Model

Introduction

This document presents model diagnostic plots to investigate the unit-level Small Area Estimation (SAE) EBLUP model that was estimated. In these plots, the direct estimates are represented by the SRE results.

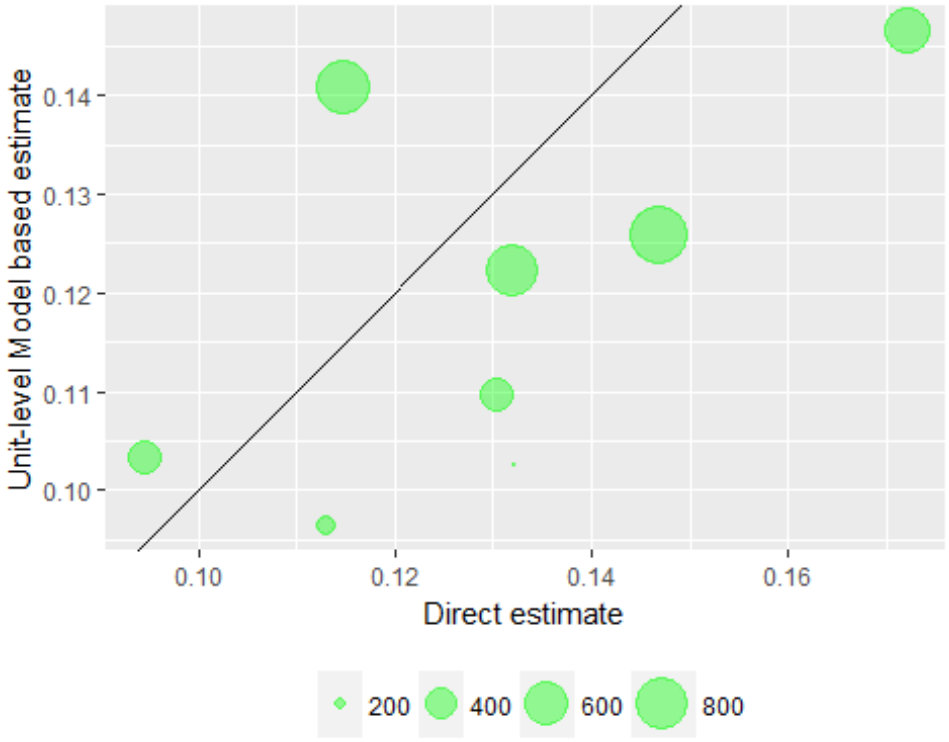
Histogram

First the histogram of the difference between unit-level model-based estimated literacy rates and the design-based estimated literacy rate (Direct) is presented. The mean (red dash line) and median (green dash line) are also presented.



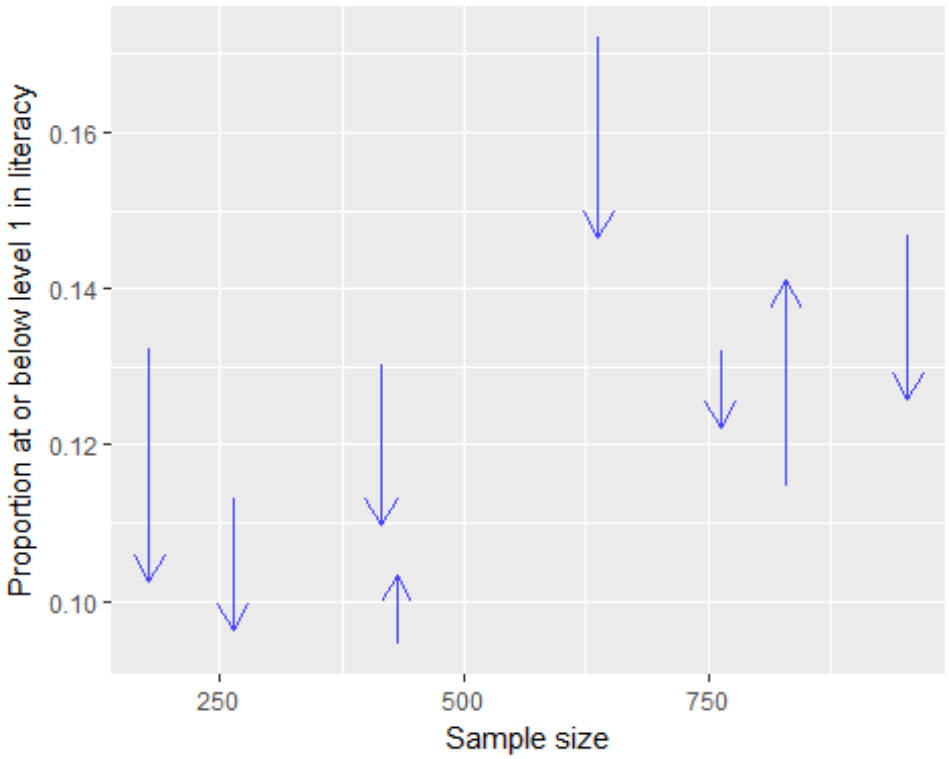
Correlation

The scatterplot shows the correlation between the unit-level model-based estimates and the direct estimates, with the direct estimates on the x-axis, and model-based estimates on the y-axis. The bubble size represent the sample sizes where larger bubbles stand for larger sizes. The black line goes across the diagonal, and it represents a perfect correlation between the two estimates.



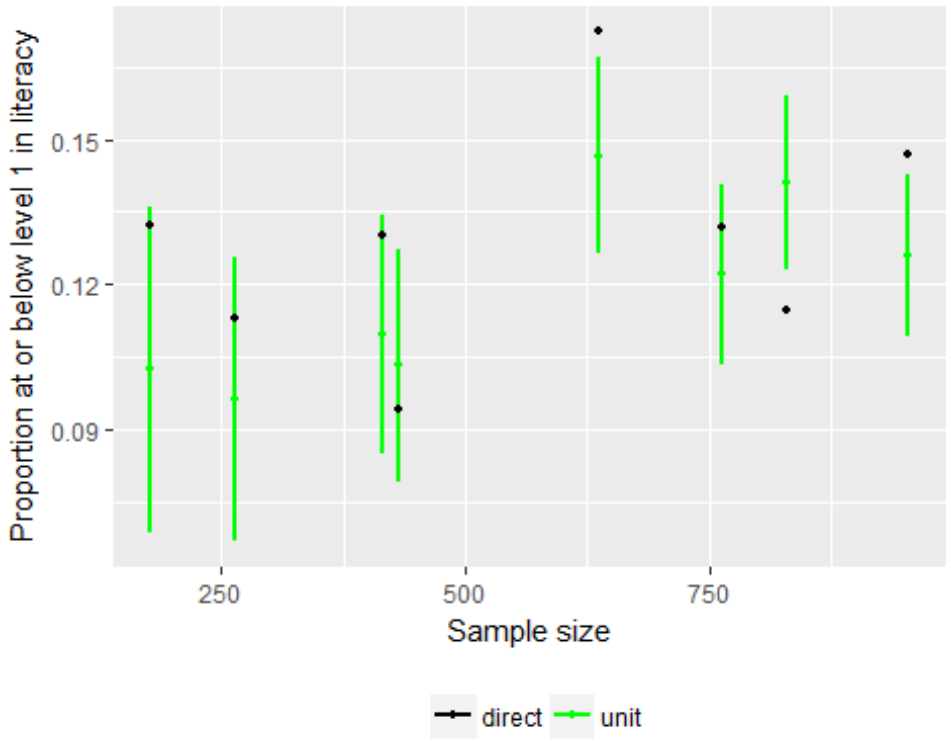
Shrinkage

The shrinkage plot shows how the models shrink the SAE estimates. The x-axis is the sample size of the small areas, and the y-axis represents the estimates. The start of the lines are the design-based estimates (Direct) and the end of the arrows are unit-level model-based small area estimates. Each model is presented separately.



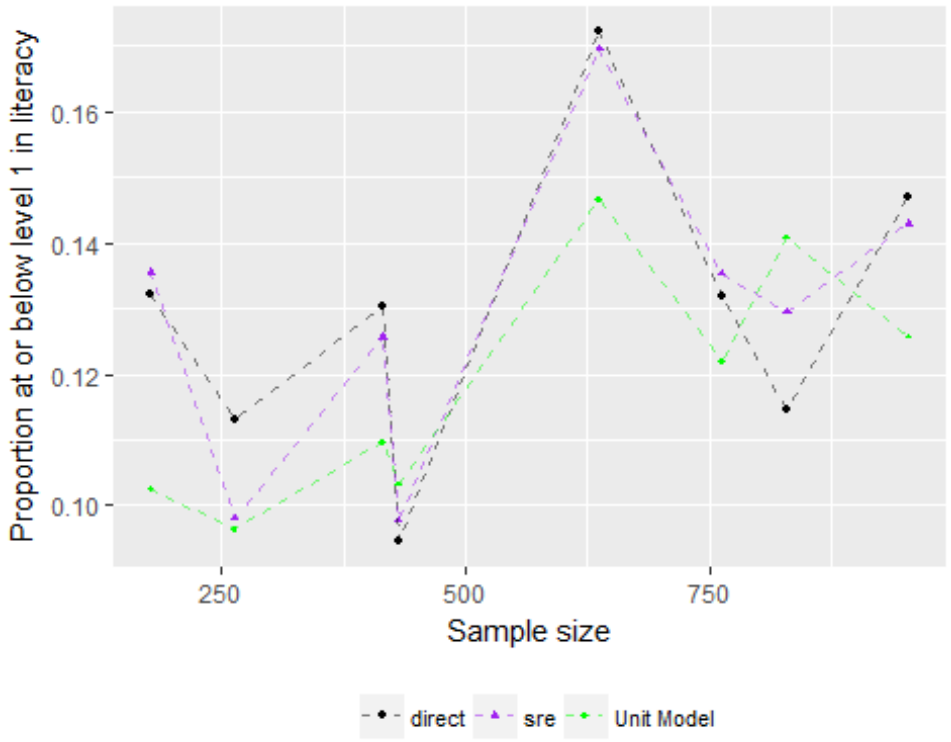
Indication of Coverage by Confidence/Credible Interval

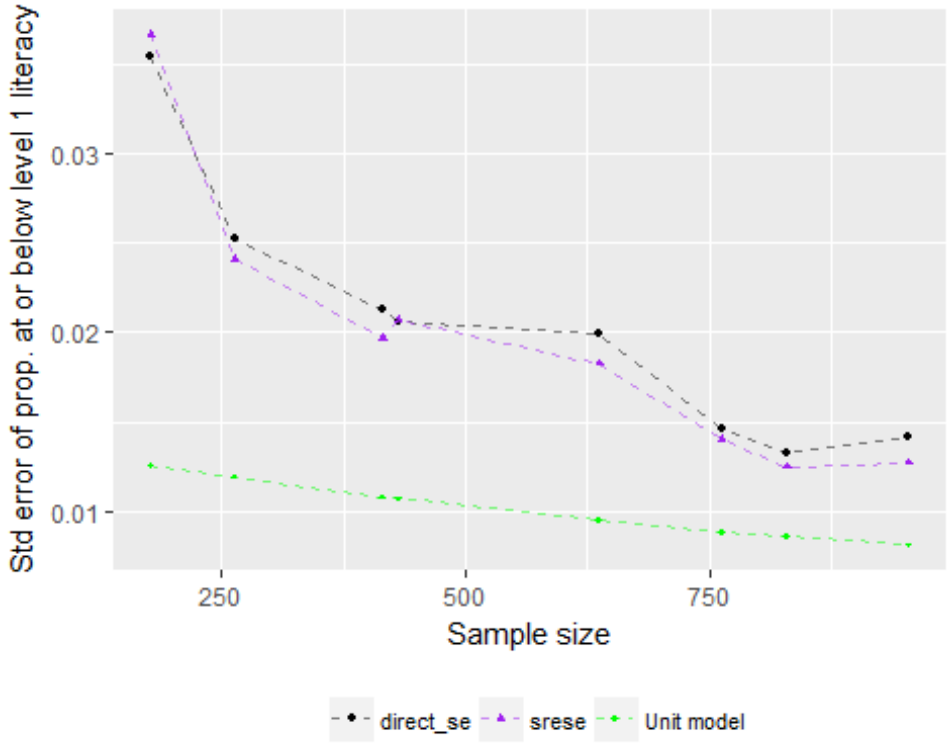
The coverage of confidence/credible interval plots present the relationship between the direct estimate and the confidence/credible interval of unit-level model estimate. The x-axis is the sample size, and y-axis is the estimates. The black dot represent the design-based estimates, and the green stands for the unit level model.



Point Estimate and Standard Error plot

The following plots show the point estimates and standard errors of the direct estimate and the unit-level model-based estimates. The x-axis is the sample size, and y-axis is the estimates. Different models can be distinguished by the color.


Proportion at or below level 1 in literacy
0.16
0.14
0.12
0.10
250
500
750
Sample size
direct
sre
Unit Model


Std error of prop. at or below level 1 literacy
0.03
0.02
0.01
250
500
750
Sample size
direct_se
srese
Unit model

Appendix C - Country Summaries of SAE Process for Estimating Proportion at or Below Level 1 in Literacy

Table C-1 Germany

Item	Description
Sampling Information	
Sample design – Cycle 1	The sample design for the 2012 PIAAC survey was comprised of a two-stage cluster sample. The first stage included a sample of 277 communities from many strata using region and urban/rural status. Controlled rounding methodology was used in the selection of the communities. The second stage included two-phases. Phase 1 involved asking communities to select and EPSEM sample of individuals from their local registry. Then in 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.
Number of completes – Cycle 1	5,465
Small areas (SAs)	Data was not provided for NUTS 2 or 3, as there are too few areas for NUTS 2 (39) and too many areas for NUTS 3 (429) to carry out the analyses reasonably. Instead data were provided for spatial planning regions (in German: Raumordnungsregionen). In total there are 96 of these regions, but a number of areas were collapsed, resulting in 85 areas overall. These regions are derived from NUTS 3 (as of December 31, 2015) and are spacious, clearly separated spatial units for federal spatial planning reporting. Spatial planning regions are used for the functional structuring of the territory of the Federal Republic of Germany for the purpose of regional planning.
Number of SAs (with sample)	85 (85) collapsed spatial planning regions
Number of SAs with number of first stage clusters within SA > 1	Unknown
Number of SAs with n > 30	73
Number of SAs with n > 100	13
Number of SAs used in models	85
Input Data	
PIAAC data received	In lieu of PIAAC microdata with an ID with the SAs, Germany provided direct estimates and standard errors for each SA.
Covariates data received	One-way tabulations of covariates were provided for each SA. Variables included gender, age groups (4 levels), nationality (4 levels), education attainment (5 levels), and employment status (3 levels).
Direct Estimation	

Item	Description
Direct estimation for point estimates	Hajek ratio estimator, as computed by Germany Variances were smoothed
Covariates used in survey regression estimator	Not applicable
Direct estimation for variances	Computed using Taylor Series and Delete-one Jackknife, as computed by Germany. Taylor Series results were used as input to the area models.
Treatment of PVs	Imputation error is addressed by using the traditional multiple imputation formula in the computation of direct estimates.
SAE Model	
Models processed	Fay-Herriot Area-level HB linear matched Area-level HB nonlinear unmatched
Covariates used in area models	Gender (1 level), age groups (3 levels), education attainment (3 levels), employment (2 levels), nationality (1 level)
Covariates used in unit model	Not applicable
SAE Benchmarking	Not applicable for this phase of research
SAE Evaluation	
Diagnostics	Limited model diagnostics performed. In the future, model diagnostics may include internal model validation, cross-validation, and the posterior predicted p-values, model sensitivity, including the use of different priors, different sets of predictor variables, and the use of the deviance information criterion (DIC) measure to compare models.
Evaluation metrics	Evaluation metrics included in this research are: Histograms of relative difference from direct estimates, by model. Bubble plots of direct estimates by each model result, with size of bubble related to sample size. Shrinkage plots with arrows showing the direction from direct estimates to model estimates, by sample size. Lack of bias plots, showing the confidence interval of the model estimate, and the direct point estimates, by sample size. MSE plots showing the resulting MSEs from direct and models, by sample size. In the future, external checks can be done, including the comparisons of aggregates to the national level with national estimates for metropolitan statistical area (MSA) status, and other characteristics based on coarsened county-level percentages of characteristics (e.g., estimated proportion in Level 1 or below in counties in the lower third of percentage with less than a high school degree).

Table C-2 Italy

Item	Description
Sampling Information	
Sample design – Cycle 1	The sample design for the 2012 PIAAC survey was comprised of a three-stage cluster sample. The first stage included a sample of 260 area primary sampling units (PSUs) selected with probabilities proportionate to size sorted by total population within explicit strata based on equal sized regions. In the second stage, a frame of dwelling units was formed from the registry. 11,592 dwelling units were selected within sampled PSUs. One person from each DU was pre-selected from the DU registry. A screener questionnaire was administered to selected DUs. If the household composition was found to be different from the registry, persons were sorted by gender and age and the selection grid is used.
Number of completes – Cycle 1	4,621
Small areas (SAs)	Provinces
Number of SAs (with sample)	110 (91)
Number of SAs with number of first stage clusters within SA > 1	71
Number of SAs with n > 30	56
Number of SAs with n > 100	6
Number of SAs used in area models	90
Coverage of SAs	Three variables (education attainment, employment status and citizenship) were each split into two levels (high, low) based on their distribution of SAs in the population. A 3-way cross-tabulation of SAs (2x2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For Italy, the coverage is 8 out of 8 cells. When excluding SAs with less than 30 cases, the coverage is 7 out of 8 cells. Improvement to the SAEs can be made by adjusting the sample design to cover all cells.
Input Data	
PIAAC data received	PIAAC microdata were provided with an ID for the SAs. All covariates mentioned below were included except citizenship and marital status.
Covariates data received	A full cross-tabulation of the following covariates was provided for each SA: Gender, single year of age, citizenship (2 levels), education attainment (6 levels), employment status (7 levels), number of people in the household (5 levels), marital status (6 levels).
Direct Estimation	
Direct estimation for point estimates	Initially, the Hajek estimator is used, with results adjusted by the survey regression estimator.
Covariates used in survey regression estimator	Gender (1 level), Age (mean), Education attainment (2 levels), employment status (1 level), number of people in the household (2 levels).
Direct estimation for variances	Computed using stratified jackknife on the residuals from the survey regression estimates. Variances were smoothed.
Treatment of PVs	Imputation error is addressed by using the traditional multiple imputation formula in the computation of direct estimates and variance estimates. That is,

Item	Description
	the Hajek and SRE are processed 10 times, and the multiple imputation formulae are then applied to arrive at the point estimate and variance estimate for each SA.
SAE Model	
Models processed	Unit-level empirical best linear unbiased prediction Fay-Herriot Area-level HB linear matched Area-level HB nonlinear unmatched
Covariates used in area models	Gender (1 level), Age (mean), Education attainment (2 levels), employment status (2 level), number of people in the household (2 levels), citizenship (1 level), marital status (1 level).
Covariates used in unit model	Gender (1 level), Age (mean), Education attainment (2 levels), employment status (1 level), number of people in the household (2 levels).
SAE Benchmarking	Not applicable for this phase of research
SAE Evaluation	
Diagnostics	Limited model diagnostics performed. In the future, model diagnostics may include internal model validation, cross-validation, and the posterior predicted p-values, model sensitivity, including the use of different priors, different sets of predictor variables, and the use of the deviance information criterion (DIC) measure to compare models.
Evaluation metrics	Evaluation metrics included in this research are: Histograms of relative difference from direct estimates, by model. Bubble plots of direct estimates by each model result, with size of bubble related to sample size. Shrinkage plots with arrows showing the direction from direct estimates to model estimates, by sample size. Lack of bias plots, showing the confidence interval of the model estimate, and the direct point estimates, by sample size. MSE plots showing the resulting MSEs from direct and models, by sample size. In the future, external checks can be done, including the comparisons of aggregates to the national level with national estimates for metropolitan statistical area (MSA) status, and other characteristics based on coarsened county-level percentages of characteristics (e.g., estimated proportion in Level 1 or below in counties in the lower third of percentage with less than a high school degree).

Table C-3 New Zealand

Item	Description
Sampling Information	
Sample design – Cycle 1	The sample design for the 2014 PIAAC survey was comprised of a four-stage cluster sample. The first stage included a sample of 1,000 area clusters (PSUs) selected with probabilities proportionate to the number of occupied dwelling units (and units under construction) sorted by total population within explicit strata based on equal sized regions. In the second stage, 1,000 meshblocks were selected using the same size measure as PSUs, one from each PSU. In the third stage, the frame of dwelling units was sorted by geography and 16,392 dwelling units were selected within sampled meshblocks. A screener questionnaire was administered to selected DUs and one person was selected per DU.
Number of completes – Cycle 1	6,177
Small areas (SAs)	Combination of Territorial Authority (TA) and Community Boards (CB)
Number of SAs (with sample)	87 comprised of 66 TAs + 21 CBs (84 comprised of 64 TAs + 20 CBs)
Number of SAs with number of first stage clusters within SA > 1	83
Number of SAs with n > 30	63 comprised of 44 TAs + 19 CBs
Number of SAs with n > 100	16 comprised of 9 TAs + 7 CBs
Number of SAs used in area models	81
Coverage of SAs	Three variables (education attainment, employment status and birthplace) were each split into two levels (high, low) based on their distribution of SAs in the population. A 3-way cross-tabulation of SAs (2x2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For New Zealand, the coverage is 8 out of 8 cells. When excluding SAs with less than 30 cases, the coverage is still 8 out of 8 cells. Improvement to the SAEs can be made by adjusting the sample design to cover cells of a 4-way table, by adding another key variable to the cross-tabulation.
Input Data	
PIAAC data received	PIAAC microdata were provided with an ID for the SAs. All covariates mentioned below were included except ethnic group.
Covariates data received	A full cross-tabulation of the following covariates was provided for each SA: Gender, age group (6 levels), birth place (2 levels), education attainment (4 levels), work and labor force status (2 levels), ethnic group (2 levels).
Direct Estimation	
Direct estimation for point estimates	Initially, the Hajek estimator is used, with results adjusted by the survey regression estimator.
Covariates used in survey regression estimator	Gender, Age Group (4 levels), Birthplace (1 level).
Direct estimation for variances	Computed using delete-one jackknife on the residuals from the survey regression estimates. Variances were smoothed.
Treatment of PVs	Imputation error is addressed by using the traditional multiple imputation

Item	Description
	formula in the computation of direct estimates and variance estimates. That is, the Hajek and SRE are processed 10 times, and the multiple imputation formulae are then applied to arrive at the point estimate and variance estimate for each SA.
SAE Model	
Models processed	Unit-level empirical best linear unbiased prediction Fay-Herriot Area-level HB linear matched Area-level HB nonlinear unmatched
Covariates used in area models	Gender, Age Group (4 levels), Education attainment (4 levels), Employment status (1 level), Birthplace (1 level), Ethnicity (2 levels).
Covariates used in unit model	Gender, Age Group (4 levels), Birthplace (1 level).
SAE Benchmarking	Not applicable for this phase of research
SAE Evaluation	
Diagnostics	Limited model diagnostics performed. In the future, model diagnostics may include internal model validation, cross-validation, and the posterior predicted p-values, model sensitivity, including the use of different priors, different sets of predictor variables, and the use of the deviance information criterion (DIC) measure to compare models.
Evaluation metrics	Evaluation metrics included in this research are: Histograms of relative difference from direct estimates, by model. Bubble plots of direct estimates by each model result, with size of bubble related to sample size. Shrinkage plots with arrows showing the direction from direct estimates to model estimates, by sample size. Lack of bias plots, showing the confidence interval of the model estimate, and the direct point estimates, by sample size. MSE plots showing the resulting MSEs from direct and models, by sample size. In the future, external checks can be done, including the comparisons of aggregates to the national level with national estimates for metropolitan statistical area (MSA) status, and other characteristics based on coarsened county-level percentages of characteristics (e.g., estimated proportion in Level 1 or below in counties in the lower third of percentage with less than a high school degree).

Table C-4 Slovakia

Item	Description
Sampling Information	
Sample design – Cycle 1	The sample design for the 2012 PIAAC survey was comprised of a two-stage cluster sample. The first stage included a sample of 562 municipalities (PSUs) selected with probabilities proportionate to the number of adults 16 to 65 years old, sorted by total population within explicit strata based on region and municipality size. In the second stage, within PSUs, persons on the population registry were sorted by gender and age and selected using a systematic random sample.
Number of completes – Cycle 1	5,723
Small areas (SAs)	Districts/counties (LAU_1)
Number of SAs (with sample)	79 (79)
Number of SAs with number of first stage clusters within SA > 1	79
Number of SAs with n > 30	64
Number of SAs with n > 100	45
Number of SAs used in area models	77
Coverage of SAs	Three variables (education attainment, employment status and nationality) were each split into two levels (high, low) based on their distribution of SAs in the population. A 3-way cross-tabulation of SAs (2x2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For Slovakia, the coverage is 8 out of 8 cells. When excluding SAs with less than 30 cases, the coverage is 7 out of 8 cells. Improvement to the SAEs can be made by adjusting the sample design to cover all cells.
Input Data	
PIAAC data received	PIAAC microdata were provided with an ID for the SAs. No covariates mentioned below were included on the small area file, however, gender, age, education and language were available on the 2012 PIAAC file.
Covariates data received	Tabulations (not full cross-tabulations) of the following covariates were provided for each SA: Gender, age group (21 levels of 5 year increments ranging from 5 to 109), nationality (16 levels), education attainment (9 levels), economic activity (13 levels), language spoken at home (14 levels).
Direct Estimation	
Direct estimation for point estimates	Initially, the Hajek estimator is used, with results adjusted by the survey regression estimator.
Covariates used in survey regression estimator	Gender (1 level), Age Group (4 levels).
Direct estimation for variances	Computed using the paired jackknife on the residuals from the survey regression estimates. Variances were smoothed.
Treatment of PVs	Imputation error is addressed by using the traditional multiple imputation formula in the computation of direct estimates and variance estimates. That is, the Hajek and SRE are processed 10 times, and the multiple imputation

Item	Description
	formulae are then applied to arrive at the point estimate and variance estimate for each SA.
SAE Model	
Models processed	Unit-level empirical best linear unbiased prediction Fay-Herriot Area-level HB linear matched Area-level HB nonlinear unmatched
Covariates used in area models	Gender (1 level), Age Group (4 levels), Education Attainment (3 levels), Employment Status (2 levels), Language (1 level), Nationality (1 level).
Covariates used in unit model	Gender (1 level), Age Group (4 levels), Education Attainment (2 levels).
SAE Benchmarking	Not applicable for this phase of research
SAE Evaluation	
Diagnostics	Limited model diagnostics performed. In the future, model diagnostics may include internal model validation, cross-validation, and the posterior predicted p-values, model sensitivity, including the use of different priors, different sets of predictor variables, and the use of the deviance information criterion (DIC) measure to compare models.
Evaluation metrics	Evaluation metrics included in this research are: Histograms of relative difference from direct estimates, by model. Bubble plots of direct estimates by each model result, with size of bubble related to sample size. Shrinkage plots with arrows showing the direction from direct estimates to model estimates, by sample size. Lack of bias plots, showing the confidence interval of the model estimate, and the direct point estimates, by sample size. MSE plots showing the resulting MSEs from direct and models, by sample size. In the future, external checks can be done, including the comparisons of aggregates to the national level with national estimates for metropolitan statistical area (MSA) status, and other characteristics based on coarsened county-level percentages of characteristics (e.g., estimated proportion in Level 1 or below in counties in the lower third of percentage with less than a high school degree).

Table C-5 Sweden

Item	Description
Sampling Information	
Sample design – Cycle 1	The sample design for the 2012 PIAAC survey was comprised of a one-stage simple random sample within explicit strata. Strata were formed from gender, age, country of birth and level of education.
Number of completes – Cycle 1	4,469
Small areas (SAs)	NUTS 2
Number of SAs (with sample)	8 (8)
Number of SAs with number of first stage clusters within SA > 1	8
Number of SAs with n > 30	8
Number of SAs with n > 100	8
Coverage of SAs	Two variables (education attainment, and birthplace) were each split into two levels (high, low) based on their distribution of SAs in the population. A 2-way cross-tabulation of SAs (2x2) was processed on both the population and sample of SAs. An indication of the coverage of the SAs is determined by the number of cells in the population that is covered by the sample. For Italy, the coverage is 4 out of 4 cells. When excluding SAs with less than 30 cases, the coverage is still 4 out of 4 cells. If more SAEs are considered in the future, improvement to the SAEs can be made by adjusting the sample design to cover cells of a 3- or 4-way table, by adding another key variable or two to the cross-tabulation.
Input Data	
PIAAC data received	PIAAC microdata from the 2012 sample were used with the ID for the SAs. All covariates listed below were on the PIAAC file.
Covariates data received	Full cross-tabulations of the following covariates were provided for each SA: Gender, age group (5 levels), birth place (2 levels), education attainment (4 levels), economic activity (13 levels), language spoken at home (14 levels).
Direct Estimation	
Direct estimation for point estimates	Initially, the Hajek estimator is used, with results adjusted by the survey regression estimator.
Covariates used in survey regression estimator	Gender (1 level), Age Group (4 levels), Birthplace (1 level).
Direct estimation for variances	Computed using the paired jackknife on the residuals from the survey regression estimates.
Treatment of PVs	Imputation error is addressed by using the traditional multiple imputation formula in the computation of direct estimates and variance estimates. That is, the Hajek and SRE are processed 10 times, and the multiple imputation formulae are then applied to arrive at the point estimate and variance estimate for each SA.
SAE Model	
Models processed	A unit-level SAE EBLUP model was generated.
Covariates used in the models	Gender (1 level), Age Group (4 levels), Birthplace (1 level).

Item	Description
SAE Benchmarking	Not applicable for this phase of research
SAE Evaluation	
Diagnostics	Limited model diagnostics performed. In the future, model diagnostics may include internal model validation, cross-validation, and the posterior predicted p-values, model sensitivity, including the use of different priors, different sets of predictor variables, and the use of the deviance information criterion (DIC) measure to compare models.
Evaluation metrics	Evaluation metrics included in this research are: Histograms of relative difference of survey regression estimator from initial direct estimates. Bubble plots of direct estimates by each model result, with size of bubble related to sample size. Shrinkage plots with arrows showing the direction from direct estimates to model estimates, by sample size. Lack of bias plots, showing the confidence interval of the model estimate, and the direct point estimates, by sample size. MSE plots showing the resulting MSEs from direct and models, by sample size. In the future, external checks can be done, including the comparisons of aggregates to the national level with national estimates for metropolitan statistical area (MSA) status, and other characteristics based on coarsened county-level percentages of characteristics (e.g., estimated proportion in Level 1 or below in counties in the lower third of percentage with less than a high school degree).