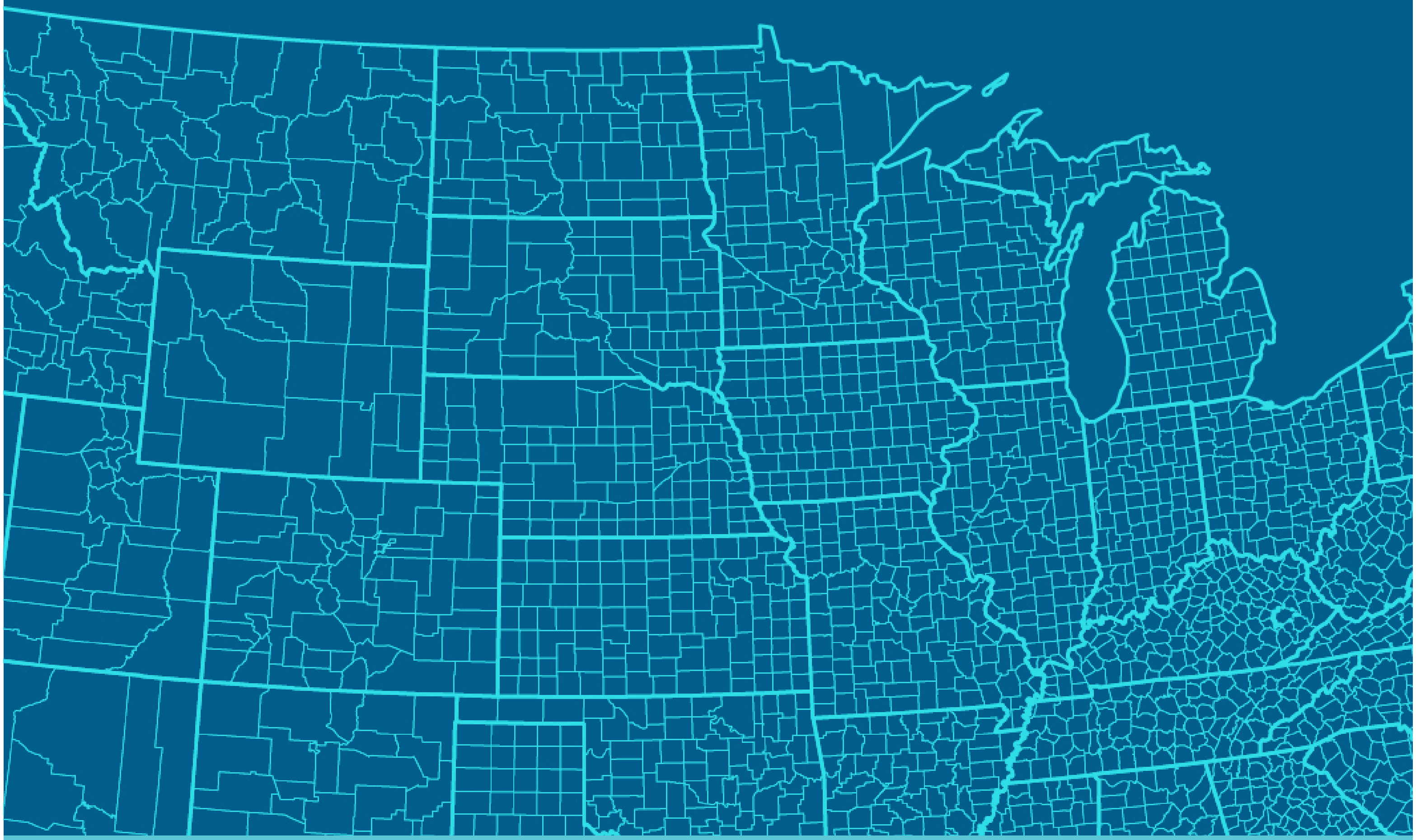
IES Institute of Education Sciences

PIAAC Small Area Estimation (SAE)

Guidance for Using State and County Estimates of Adult Skills



PIAAC Small Area Estimation (SAE): Guidance on Use of State and County Estimates

This guide explains reporting practices and statistical methods that are necessary to be accurate when using data available through the U.S. Program for the International Assessment of Adult Competencies (PIAAC) Skills Map.[1] This guide provides a brief background on these data and then provides five guidelines for (a) reporting percentages and averages and (b) conducting exploratory, descriptive, and regression analyses. For each guideline, examples and discussion points are provided.

What is the purpose of this User's Guide for the PIAAC Skills Map?

- To provide guidelines and examples for those who want to use the data from the PIAAC Skills Map to support analyses or policy recommendations.
- As the data from the PIAAC Skills Map are model-based skills estimates, they require great care to accurately analyze and report.

The U.S. Program for the International Assessment of Adult Competencies (PIAAC) Skills Map provides model-based state and county statistics of literacy and numeracy proficiency for adults 16 to 74 years old. These statistics were developed using advanced small area estimation (SAE) methods using survey data from the first cycle (2012/2014/2017) of PIAAC combined with auxiliary data from the American Community Survey (ACS 2013-2017)[2]. SAE includes a variety of statistical techniques to estimate survey outcomes for small areas. Small areas are defined as population subgroups, like geographical areas or socioeconomic groups, with insufficient survey data to produce reliable estimates. The SAE methodology allows for reliable estimates to be produced even when insufficient survey data exists.

[1] https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020047

[2] The U.S. PIAAC State and County estimates are based on area-level hierarchical Bayes linear three-fold models. Model-based estimates for the percentages *at or below Level 1* and *at or above Level 3* were constructed using a bivariate model fitted to the survey estimates for these two quantities. The model results were then used to derive the model-based percentage *at Level 2*. A univariate model was used for the average. See more details in the methodology report available at https://nces.ed.gov/pubs2020/2020225.pdf.

The PIAAC Skills Map provides average proficiency scores for all U.S. states and counties along with the percentage of the adult population at specific levels of proficiency in those states and counties. In addition to displaying these various proficiency estimates, the Skills Map online tool provides each point estimate's associated measures of uncertainty in terms of credible intervals[3] (at 95%) and coefficients of variation (CVs). The Skills Map's website data (PIAAC SAE data) available for download include the lower and upper bounds of the credible intervals and the CVs. The standard error associated with a point estimate can be derived as the product of the point estimate and the CV[4]. The variance estimates are the standard errors squared.

The PIAAC SAE data allows interested researchers to conduct various statistical analyses that combine the model-based skills estimates with data from other sources. This can be done by linking the PIAAC SAE data with other data using the county and/or state Federal Information Processing

Why is it important to include variance associated with PIAAC SAE data in analyses? If not included...

- Analysis results with calculated statistical significance may be misleading due to the false impression of precision in the results when the variance from the SAE data is not included in the analysis results.
- One may commit Type I errors.

[3] In Bayesian statistics, a credible interval is an interval for which there is a specified probability that the true value is within it, given initial assumptions of what the value may be and information provided in the data. For the PIAAC SAE, the 95 percent credible intervals were computed by calculating the 2.5 percent (lower bound) and 97.5 percent (upper bound) quantiles of the posterior distributions of county-, state-, or nation-level quantities of interest. As a result of the model fit, the posterior distributions of the quantities of interest were approximated by the 4,500 Markov chain Monte Carlo simulated samples. Credible intervals are analogous to confidence intervals in frequentist statistics. Credible intervals and confidence intervals are different in their philosophies: credible intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas the confidence intervals treat their bounds as random and the parameter as fixed. In addition, credible intervals use knowledge of the prior distribution, while the confidence intervals do not. For more details about uncertainty intervals, see Chapter 9 in Casella and Berger (2002).

[4] $SE = \hat{\theta} \times CV$, where SE denotes the standard error, $\hat{\theta}$ denotes the point estimate, and CV is the coefficient of variation.

Standards[5] (FIPS) codes. The PIAAC SAE data workbook includes one spreadsheet for **county**-level information and one for **state**-level information. The information in these spreadsheets includes FIPS codes and estimates with associated measures of uncertainty. Analyses using these data need to take into account that all model-based skills estimates are subject to error. Therefore, except for exploratory analyses[6], **the model-based skills estimates should not be treated as if they were fixed and known quantities and their associated measures of uncertainty should be included in reports or analyses using these estimates.**

This document provides general guidance on reporting and conducting analyses using the PIAAC SAE data. Guidelines are provided for reporting percentages and averages and for conducting exploratory, descriptive, and regression analyses. Examples are provided for each guideline, often followed by a discussion. These guidelines are by no means comprehensive, as users may be interested in different analyses using the PIAAC SAE data than the ones provided here.

What can PIAAC SAE data support?

- Reporting
- Research analysis
  - Exploratory
  - Descriptive
  - Regression

Five guidelines provide examples and discussions to assist data users in these efforts.

The following five guidelines cover a range of applications using the PIAAC SAE data, from simple reporting to complex statistical analyses. Guidelines 1 through 3 are for reporting the PIAAC model-based skills estimates for literacy or numeracy. The examples in these three guidelines involve minimal but necessary computations. Basic statistical knowledge is assumed for the first two guidelines, and intermediate statistical knowledge is assumed for the third guideline. Guidelines 4 and 5 extend beyond the reporting of the PIAAC model-based skills estimates and dive into research

[5] For more information, see https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt#:~:text=FIPS%20codes%20are%20numbers%20which,to%20which%20the%20county%20belongs.

[6] See, for example, the classification tree example in Guideline 5.

analyses. These guidelines are intended for users with more advanced statistical knowledge.

1. Guideline 1 – Reporting model-based skills estimates 5

2. Guideline 2 – Estimating a skills population total using model-based estimates 5

3. Guideline 3 – Estimating skills percentages or averages for groups of counties or states using model-based skills estimates 7

4. Guideline 4 – Relating the model-based skills estimates to data from external sources 9

5. Guideline 5 – Using contextual variables to describe groups of counties or states...14

References 17

What analyses using the PIAAC SAE data are not appropriate?

- Regression analysis with predictors including the model-based skills estimates and values of one or more variables used as predictors in the PIAAC SAE models will lead to confounding effects and misleading regression coefficients.
- Regression analysis with predictors including the model-based skills estimates without the associated errors will lead to misleading regression coefficients.

Note: The PIAAC SAE model predictors are: Percentage of population age 25 and over with less than high school education; Percentage of population age 25 and over with more than high school education; Percentage of population below 100 percent of the poverty line; Percentage of Black or African American population; Percentage of Hispanic population; Percentage of civilian noninstitutionalized population who has no health insurance coverage; Percentage of population age 16 and over with service occupations.

1. Guideline 1 – Reporting model-based skills estimates

When reporting PIAAC model-based skills estimates of average scale scores or percentages at proficiency levels, the associated credible intervals should be included in all reports.

> **Example 1.** For reporting a PIAAC model-based estimate of percentage, the statement should call out the credible interval: e.g., "*The percentage at or below Level 1 for literacy in Wood County, Ohio is estimated at 13%, with a 95% probability that the actual (true, unknown) percentage is between 9% and 16%.*" The variable names used in this example are: LIT_P1: the percentage at or below Level 1 for literacy; and LIT_P1_CI_L and LIT_P1_CI_U for the lower and upper 95% credible interval bounds, respectively.
>
> **Example 2.** For displaying heat maps of PIAAC model-based skills estimates extracted from the Skills Map, a statement should be given that calls out the associated error: e.g., "*The heat map displays a set of estimates that have associated uncertainty. Consult the Skills Map online tool to see estimates of low, moderate, and high confidence on the map. A dynamic tool is also available to better understand the uncertainty associated with the heat map.*"

2. Guideline 2 – Estimating a skills population total using model-based estimates

To derive estimates of skills population totals of individuals with competency scores in a certain range at the county level, the county-level PIAAC model-based skills estimates and their associated credible intervals can be multiplied by the appropriate ACS 2013-2017 county-level population total. All the necessary data for such calculations are available in the PIAAC SAE data. A similar approach can be adopted to estimate state-level skills population totals.

> **Example: Number of low-skilled adults in Los Angeles County, CA.** The PIAAC SAE data for Los Angeles County provide:
>
> a) The ACS 2013-2017 population total of individuals 16-74 years old (POP_TARGET = 7,541,627);
>
> b) The estimated percentage of "low-skilled" adults, those at or below Level 1 (LIT_P1 = 34.3%); and

c) The credible interval (at 95%) for the estimated percentage in (b) from a lower bound of 31.1% (LIT_P1_CI_L) to an upper bound of 37.6% (LIT_P1_CI_U).

To compute the estimated population total of adults 16 to 74 years old with literacy scores at or below Level 1 in Los Angeles County, multiply the ACS population total in (a) by the estimated percentage of low-skilled adults in (b), i.e., 7,541,627 × 0.343 = 2,586,778. The credible interval (at 95%) associated with this estimated total can be computed by multiplying the ACS population total in (a) by the credible interval bounds in (c), i.e., the lower bound is 7,541,627 × 0.311 = 2,345,446, and the upper bound is 7,541,627 × 0.376 = 2,835,652.

Discussion 1. This approximation assumes that the ACS county-level population total has no associated error. However, the ACS county-level population totals are estimates and are subject to sampling error, as noted on the ACS website about sample size definitions[7]. A note should be added accordingly, such as "*The actual width of the credible interval constructed above might be larger than what is reported because the uncertainty associated with the ACS population total is not accounted for in the computation*". Such a note would apply to all counties, but would be especially important for counties with a small ACS sample size. This is because the sampling errors in the ACS county-level population totals for small counties might be larger than for larger counties.

Discussion 2. Using this approach assumes that the skills estimates are representative of the same population measured by the ACS. Although a population adjustment[8] was made to the PIAAC survey estimates used as input into the SAE models, the underlying PIAAC survey weights were not fully calibrated to the ACS population totals. However, an evaluation of aggregated estimates shows close alignment between aggregated estimates and survey estimates at the national level, where the survey estimates are stable. See Krenzke et al. (2020), Section 6.2, for details.

[7] https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/sample-size-definitions.html; see, for example, items 2 and 3 on this website.

[8] A population adjustment was implemented using a survey regression approach. Two purposes of the survey regression adjustment were (1) to reduce the mean square error of the PIAAC survey estimates, and (2) to more closely align the PIAAC sample distribution to the ACS population distribution; see Krenzke et al. (2020) for more details.

3. Guideline 3 – Estimating skills percentages or averages for groups of counties or states using model-based skills estimates

To derive skills estimates for groups of counties or states, weighted averages can be computed with weights given by the ACS 2013-2017 population totals of individuals 16 to 74 years old. Associated uncertainty measures require approximations. These statistical calculations are more involved and require cautionary notes (see the discussion below).

Example: Average numeracy score for Pacific Northwest states. The PIAAC SAE data for Idaho, Oregon, and Washington provide:

a) The ACS 2013-2017 population totals of individuals 16 to 74 years old in each state (POP_TARGET),

b) The estimated average numeracy score for each state (NUM_A), and

c) The credible intervals (at 95%) for the estimated average scores in (b) (the lower bound of the 95% credible interval for average numeracy is NUM_A_CI_L, and the upper bound is NUM_A_CI_U).

To compute the estimated average numeracy score for these three states combined, take the weighted average of the three state-level estimated average numeracy scores in (b), using the corresponding ACS population totals in (a) as aggregation weights. That is, sum the product of the ACS population totals and the average numeracy scores, and then divide by the sum of the ACS population totals:

$$\frac{[\overset{\text{(Idaho)}}{(1{,}174{,}765 \times 256.4)} + \overset{\text{(Oregon)}}{(2{,}996{,}723 \times 260.4)} + \overset{\text{(Washington)}}{(5{,}323{,}955 \times 262.4)}]}{1{,}174{,}765 + 2{,}996{,}723 + 5{,}323{,}955} = 261.0$$

To compute the estimated variance of the group skills estimate, sum the product of the ACS population totals and the skills estimated variances[9] for the counties (or states) in the group, and then divide by the sum of the ACS population totals

[9] The skills estimated variance is the posterior variance of the target quantity (e.g., average numeracy score), i.e., the simple variance of the estimated target quantity across the 4,500 Markov chain Monte Carlo samples that simulate the posterior distribution of the target quantity.

for the counties (or states) in the group. Continuing with the example, the estimated variance of the group estimated average numeracy score is computed as:

$$\frac{[\overset{\text{(Idaho)}}{(1{,}174{,}765 \times 12.89)} + \overset{\text{(Oregon)}}{(2{,}996{,}723 \times 13.29)} + \overset{\text{(Washington)}}{(5{,}323{,}955 \times 9.91)}]}{1{,}174{,}765 + 2{,}996{,}723 + 5{,}323{,}955} = 11.35$$

As noted before, the skills estimated variances are not directly available from the PIAAC SAE data. However, they can be derived as the squares of the products between the skills point estimates and their associated CVs. For example, for Idaho, the skills estimated variance of the estimated average numeracy score is equal to $(256.4 \times 0.014)^2 = (3.59)^2 = 12.89$.

As shown above, the estimated variance of the estimated average numeracy score for the group of Pacific Northwest states is equal to 11.35. The square root of this quantity, 3.37, is the standard error of the estimated group average numeracy score. This group standard error multiplied by 1.96 is an approximated margin of error (95% confidence interval half-width) associated with the estimated group average numeracy score, i.e., 1.96 × 3.37 = 6.60. Apply the margin of error to the estimated group average numeracy score, by adding and subtracting it. The result is that there is 95% confidence that the average numeracy score for the group of Pacific Northwest states is between 254.4 (=261.0-6.6) and 267.6 (=261+6.6).

Discussion 1. In the approach illustrated above, the standard error associated with the skills estimate for groups of counties (or states) is approximated by treating the counties (or states) as independent. However, this assumption does not hold in practice. Therefore, the derived estimated variance of the estimated group average numeracy score is biased.

Also, while credible intervals are asymmetric, this approximation will result in a symmetric interval. The symmetric interval is more likely to include zeros when estimating a small score. As a consequence, one could claim there are fewer significantly different from zero quantities than when credible intervals would be analyzed.

To obtain better variance estimates and associated uncertainty intervals, one should use the posterior samples for the parameters involved in the grouping.

These samples are obtained by fitting and estimating the PIAAC SAE models. Using these model results, one can approximate the posterior distribution of the group quantity of interest. Posterior summaries, such as variance and credible intervals, could then be constructed using the approximated group posterior distribution. Since general users do not have access to these samples, we proposed the above naïve approach based on additional assumptions. These samples are available by request[10].

Discussion 2. Aggregating skills estimates using this approach assumes that the ACS population totals have no associated error and the skills estimates are representative of the same population measured by the ACS. See Discussions 1 and 2 of Guideline 2 for more detail.

4. Guideline 4 – Relating the model-based skills estimates to data from external sources

This guideline emphasizes the need for incorporating the uncertainty related to the model-based skills estimates in all computations that include the model-based estimates. A regression analysis using the model-based skills estimates as observations of predictors subject to error is provided as an illustrative example. We adopt a Bayesian inferential approach in the example, but the user is free to use a frequentist approach if so desired; for details on Bayesian analysis see Gelman et al. (2013) available at http://www.stat.columbia.edu/~gelman/book/BDA3.pdf.

Example: Regression analysis using the PIAAC model-based skills estimates as observations of predictors

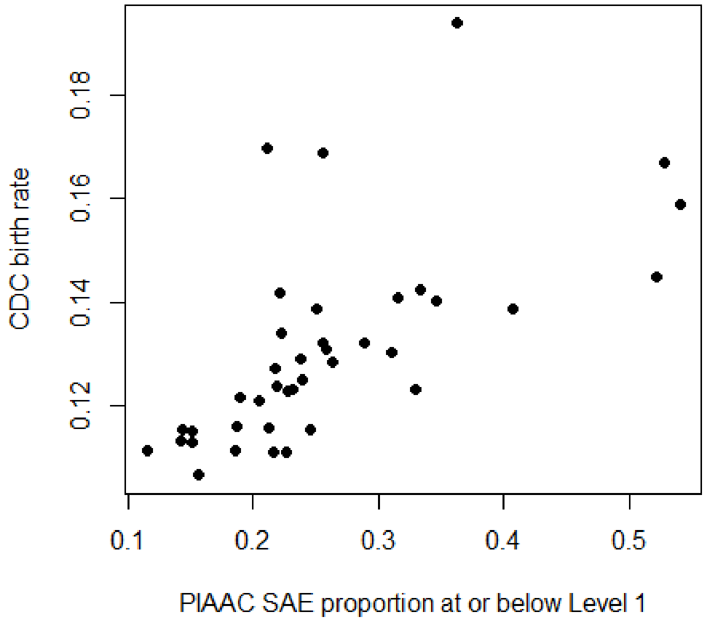
Data preparation. A first step for such analyses is to link the proficiency estimates available in the PIAAC SAE data with data from other sources. This linking can be achieved using the county and/or state FIPS codes, provided that the external data contains such information. To illustrate, birth rate data were downloaded from https://wonder.cdc.gov/natality-expanded-current.html on May 25, 2021. In particular, a county-level data file containing the 2016-2019 birth rates in Texas, with option 'all' for all the categories required, was extracted from this website. The county-level birth rates data for individual counties were merged with the county-level literacy estimates available in the

[10] Available upon request Holly.xie@ed.gov

PIAAC SAE data using county FIPS codes. We note that birth rates for counties with population below 100,000 were published for a collapsed category rather than for individual counties. In this example, we only use the data for 39 counties in Texas with population at least 100,000.

Exploratory data analysis. As a second step, prior to model fit and estimation, a scatterplot was produced as part of exploratory data analysis; see Figure 1. The scatterplot in Figure 1 reveals a positive association between the county-level birth rates and the county-level proportions at or below Level 1 literacy.

Figure 1. Scatterplot of the county-level birth rates by the county-level proportions at or below Level 1 literacy in Texas counties with population at least 100,000



NOTE: The birth rates data were downloaded from https://wonder.cdc.gov/natality-expanded-current.html on May 25, 2021. The PIAAC SAE proportions at or below Level 1 literacy (LIT_P1) are model-based estimates available from the PIAAC SAE data.

Model specification. Next, a heteroskedastic simple linear regression may be fitted to the county-level data. The response variable would be the birth rate, with observations available at the county-level and denoted by y_i, where i is an index for county. The predictor variable would be the proportion at or below Level 1 in literacy (LIT_P1), with observations available at the county-level and denoted by x_i^*. The observations x_i^* are model-based measurements of the true underlying proportions at or below Level 1, observed with associated variance estimates $\sigma^2_{u,i}$. That is, the pairs of observations $(x_i^*, \sigma^2_{u,i})$ are the skills point estimates and their associated variances available from the PIAAC SAE data. Recall that the variance estimates are not directly available, but could be derived from the point estimates (LIT_P1) and the CVs (LIT_P1_CV).

Since the predictor variable is random, and its county-level observations are subject to county-specific error, the simple linear regression model must include a measurement error component. The model would be specified as follows:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\ x_i^* &= x_i + u_i, \quad u_i \sim N(0, \sigma^2_{u,i}) \\ x_i &= \mu + v_i, \quad v_i \sim N(0, \sigma^2_v) \end{aligned}$$

Let this model be denoted by LMS (for linear model with structural measurement error). The regression parameters β_0 and β_1 are unknown, and the random errors ϵ_i are assumed to have mean 0 and variance σ^2. The mean parameter μ and the variance parameters σ^2 and σ^2_v are also unknown. Note that LMS is specified as a multilevel model, and the unobserved observations x_i are used in the first level. Instead of observing the true proportions at or below Level 1 in literacy, x_i, we observe their PIAAC model-based skills estimates x_i^*.

For comparison, in this example, we are considering the LMS model as well as a simple linear regression model that ignores the error in the predictor variable. Let this model be denoted by LM (for linear model) and be specified as follows:

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Once again, the regression parameters β_0 and β_1 and the variance parameter σ^2 are unknown. Note that, unlike in the LMS, the PIAAC model-based skills estimates x_i^* are assumed to be fixed and known in the LM.

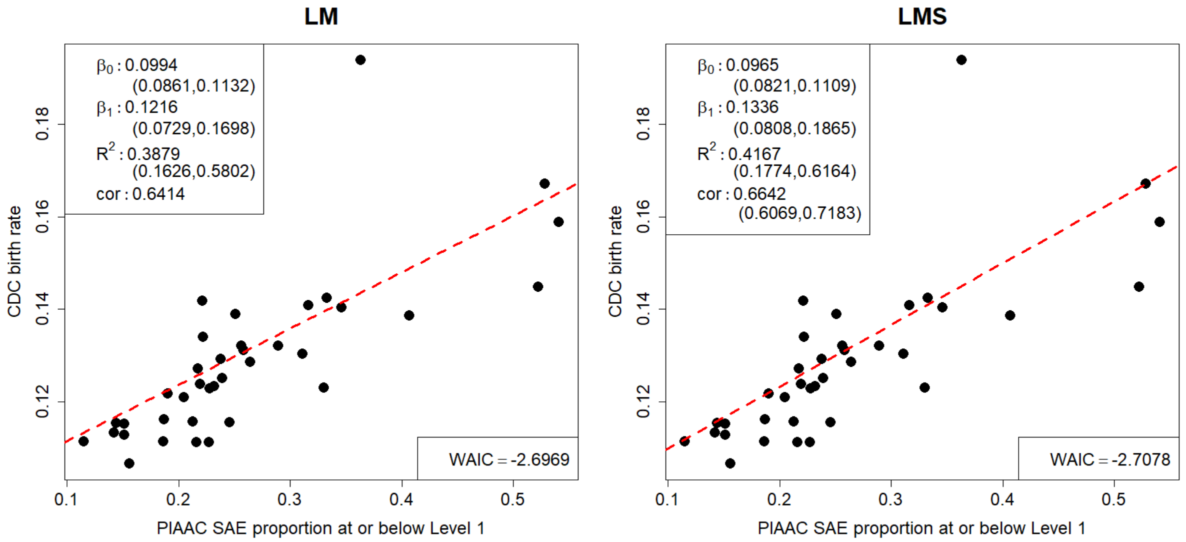
Model fit, validation, and estimation. Model fit and estimation may vary across users. In this example, we adopt a full hierarchical Bayes specification for models LM and LMS and fit the models using Markov chain Monte Carlo (MCMC). For this, weakly informative prior distributions are adopted for the model parameters, and an MCMC configuration specified. Model diagnostics, including mixing and convergence for the MCMC samples and residual analyses, are conducted to ensure that valid inferences may be conducted. The details are not included here.

Posterior distributions are produced for all model parameters and approximated by a large number of MCMC samples. Using these model results, we define a Bayesian correlation for the LMS model as a measure of association between the response and the predictor variables. For the LM model, the correlation between the response and the predictor is simply the sample correlation, since the observations for the predictor variables are assumed fixed and known. Also, we define a Bayesian R^2 for both the LM and the LMS models to estimate the proportion of the variance in the response variable that is predictable from the predictor variable using the linear regression assumed between these two variables. Once again, the details are not included here.

Model results and comparison. Finally, we compare the LMS and the LM models fits using the Widely Applicable Information Criterion (WAIC); see Watanabe (2013) available at https://www.jmlr.org/papers/volume14/watanabe13a/watanabe13a.pdf. The WAIC values, as well as the posterior means and 95% credible intervals for a) the regression parameters (β_0 and β_1), b) the association metric (cor), and c) the determination metric (R^2), are provided in Figure 2 legends. For each parameter, the 95% credible intervals are provided in parentheses under the posterior means, as applicable. The best fitted regression lines are illustrated in red in the scatterplots provided in Figure 2. Note that the effect of accounting for the error in the predictor variable is only mildly observed in these results:

- The estimated regression intercept is smaller under the LMS model than under the LM model. However, there is not enough information to conclude they are significantly different, because their corresponding 95% credible intervals overlap.
- The estimated regression slope, coefficient of determination, and coefficient of correlation are larger under the LMS model than under the LM model. However, there is not enough information to conclude they are significantly different, because the 95% credible intervals for the regression slope and the coefficient of determination overlap, and the correlation coefficient under the LM model is contained inside the 95% credible interval for the correlation coefficient under the LMS model.
- The LMS and the LM models have similar goodness of fit to the data, as indicated by their similar WAIC, with only a tiny improvement observed for the LMS model (smaller WAIC).

Figure 2. Scatterplots of the county-level birth rates by the county-level proportions at or below Level 1 literacy in Texas counties with population at least 100,000, along with modeling results



Discussion 1. Let r represent the estimated correlation coefficient between variables X and Y. In a simple linear regression with Y being the outcome variable and X being the predictor variable, a common assumption is that X is fixed and has no associated error. The estimated regression slope is a function of r, and therefore will be biased when X has associated error (e.g., when the observations of X are model-based skills estimates). Therefore, when using the model-based skills estimates as observations of fixed covariates in a simple linear regression model, or when computing pairwise correlations, the measures of association and determination will be biased and misleading. The illustrative example shows an approach on how to reduce that bias. The magnitude of the bias varies from study to study, so the user should not make general conclusions based on the illustrative example provided above.

Discussion 2. The illustrative example in this guideline uses data from an external source, however, the PIAAC SAE data may also be used to conduct similar analyses of interest. For example, the ACS Supplemental Nutrition Assistance Program (SNAP) data are available (variable name is SNAP) in the PIAAC SAE data and were not used to construct the PIAAC model-based skills estimates. Users interested in the relationship between the county-level ACS SNAP values and the county-level PIAAC model-based skills estimates may conduct similar analyses to the one illustrated above.

Discussion 3. The illustrative example in this guideline assumes the same distribution for all the county-level observations on the response variable. In general, this may not always hold in practice. Users should carefully test the model assumptions and account for all the necessary error sources.

5. Guideline 5 – Using contextual variables to describe groups of counties or states

This guideline is for defining groups of counties or states in terms of the model-based skills estimates. The groups of counties or states are then described in terms of other contextual variables. The purpose is to gain more information, or a picture, about the association between contextual variables and the model-based skills estimates. Users may be interested in describing counties or states in terms of other variables to gain a better understanding of the association between the model-based skills estimates and observations of the other variables. Defining groups of counties or states can be done in different ways using the PIAAC model-based skills estimates. Two examples are provided to describe, or explore, relationships between groups of PIAAC model-based skills estimates and groups of estimates from external sources:

1. Groups of counties are formed using the estimated median percentage at or below Level 1 literacy across counties, and

2. Groups of counties are formed using a classification tree analysis.

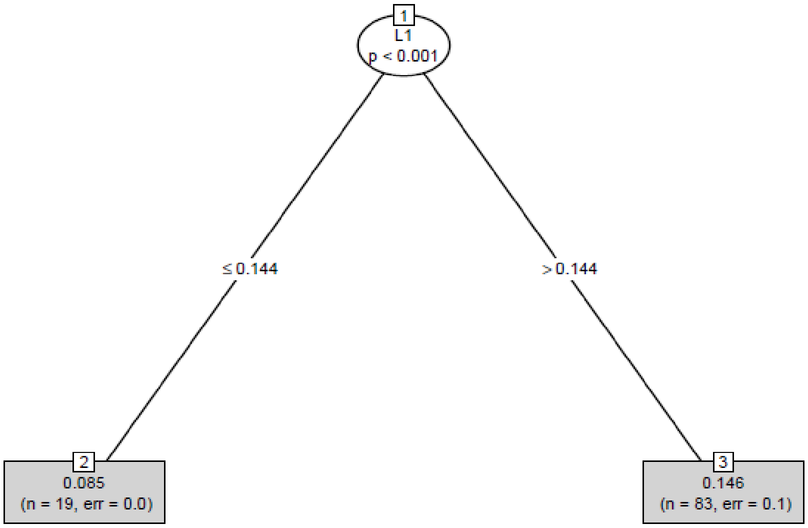
Example 1: Descriptive analysis for Illinois counties above and below median proficiency in terms of SNAP. The PIAAC SAE data for all Illinois counties provide their a) percentages at or below Level 1 in literacy, and b) percentages of households receiving SNAP benefits from the ACS 2013-2017. One may split the counties in Illinois into two groups based on the estimated median percentage at or below Level 1 in literacy across all the counties in Illinois. Among the 102 counties in Illinois, and treating each county equally, the estimated median percentage at or below Level 1 in literacy is 17.0%. For the 51 counties below the median, indicating higher literacy, the average county percentage of SNAP households is 11.3%, and for the 51 counties at or above the median, it is 15.6%.

Discussion. Note that, unlike the previous guidelines and examples, measures of uncertainty related to various estimates are not accounted for in this guideline and related examples. Here, we provide an example of an exploratory or descriptive analysis for the purpose of gaining a better understanding of the relationship between variables rather than inferential, predictive, or causal analyses. However, any conclusions must be made with care.

Example 2. Exploratory analysis using classification trees for Illinois counties in terms of SNAP. Exploring the variable of interest by grouping counties based on the model-based skills county-level estimates can be very informative. Doing so can provide a general picture of the associations between variables. Example 1 used the median to group the counties. In this example, the goal is to describe the percentage of households receiving SNAP benefits for groups of counties defined in terms of literacy. This example uses a classification tree[11] to form groups of counties. The classification tree is produced using the PIAAC model-based skills estimates in the first stage. The approach initially causes two or more branches to form, based on an analytically determined threshold of 14.4% (as shown as a proportion in Figure 3) for the estimated literacy percentage at or below Level 1. The branches each lead to a group of counties. Note that there are 19 counties with an estimated literacy percentage at or below Level 1 that is less than or equal to 14.4%. For these 19 counties, the average percentage of households receiving SNAP benefits is 8.5%. The remaining 83 counties have an estimated literacy percentage at or below Level 1 that is greater than 14.4%. For these 83 counties, the average percentage of households receiving SNAP benefits is 14.6%.

[11] For this illustration we focus on a conditional inference classification tree analysis (Ctree), as provided in Hothorn, Hornik, and Zeileis, A. (2006), using the R package *partykit* (Hothorn and Zeileis, 2015) for Kansas.

Figure 3. Classification tree analysis for Illinois data



Note: The SNAP proportions reported in the nodes of the trees are simple averages (not weighted by the population totals) over the SNAP proportions in all the counties classified as belonging to the corresponding group; for example 0.085 represents the average SNAP proportion across the 19 counties, and 0.146 represents the average SNAP proportion across the 83 counties.

Discussion 1. This approach provides a way to help describe the differences in the observations of the dependent variable by dividing into groups based on the PIAAC model-based skills estimates. The classification tree does not account for the error associated with the model-based skills estimates when grouping the counties by the model-based skills estimates. Therefore, the categorization into groups of counties may not be statistically significant. Nevertheless, this method can be used to assist with exploring associations between variables.

Discussion 2. If other variables in addition to the model-based skills estimates are of interest to form groups, then it is recommended to conduct a two-stage analysis. In the first stage, as shown above, determine if the association between the proficiency variable and the dependent variable, as assessed from their corresponding observations, can distinguish between groups of counties. In the second stage, add other covariates to determine their effect, given the proficiency variable in the model from the first stage. A two-stage approach as explained above avoids confounding effects from the use of SAE model predictors at the same time as using the model-based skills estimates.

References

Casella, G. and Berger, R.L. (2002) Statistical Inference. 2nd Edition, Duxbury Press, Pacific Grove.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013). *Bayesian Data Analysis Third edition*. Electronic edition with errors fixed as of 15 February 2021. https://users.aalto.fi/~ave/BDA3 (columbia.edu).

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, **15**(3), 651-674. doi: 10.1198/106186006X133933.

Hothorn, T. and Zeileis, A. (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research*, **16**, 3905-3909. http://jmlr.org/papers/v16/hothorn15a.html.

Krenzke, T., Mohadjer, L., Li, J., Erciulescu, A., Fay, R., Ren, W., Van de Kerckhove, W., Li, L., and Rao, J.N.K. (2020). *Program for the International Assessment of Adult Competencies (PIAAC): State and County Estimation Methodology Report* (NCES 2020-225). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Watanabe, S. (2013). A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*. 14 (2013) 867-897.