Westat

Improving Lives Through Research®

Salvaging Data from an Incomplete Sample Through Statistical Data Integration

Wendy Van de Kerckhove, Tom Krenzke, Benjamin Schneider

WESTAT @ FCSM 2024

Agenda

- Background
  - Statistical problem/motivation
  - U.S. PIAAC Cycle II
- Method for combining samples
- Evaluation/outcomes
  - National estimates
  - Models
- Summary and conclusion

Background

Statistical problem as motivation

- **Incomplete sample** | Probability sample but with unexpected disruptions
- Example | Data collection was halted during the pandemic
- Resulting sample is the outcome of:
  - Probability selection,
  - Non-probabilistic mechanism that determined which sampled cases and areas (PSUs) were worked, and
  - Nonresponse
- Goal | Salvage data from the incomplete sample
- **Solution** | Use techniques for integrating probability and non-probability samples

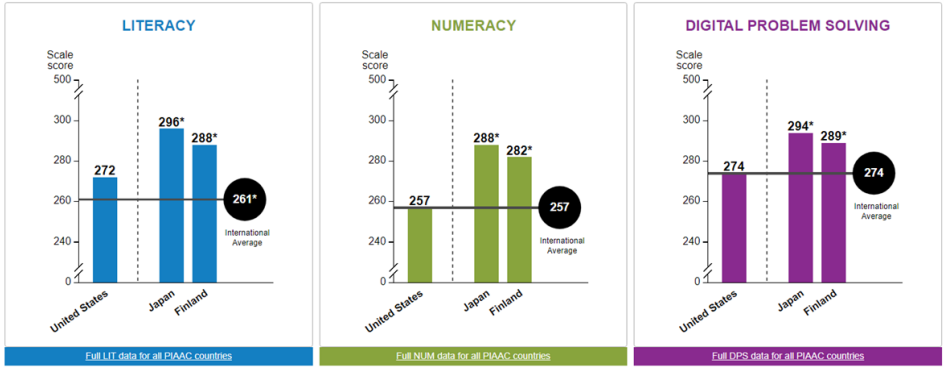
U.S. PIAAC Cyle II | Overview

- PIAAC = Program for the International Assessment of Adult Competencies
- In-person survey of adult literacy, sponsored by the National Center for Education Statistics (NCES)
- Part of international survey
- **Goal** | High quality data and sufficient yield for:
  - National estimates,
  - Item Response Theory (IRT) models, and
  - Small-area estimates for state and county (by group)

Examples from Cycle I

National estimates Overall, by domain, international comparisons

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



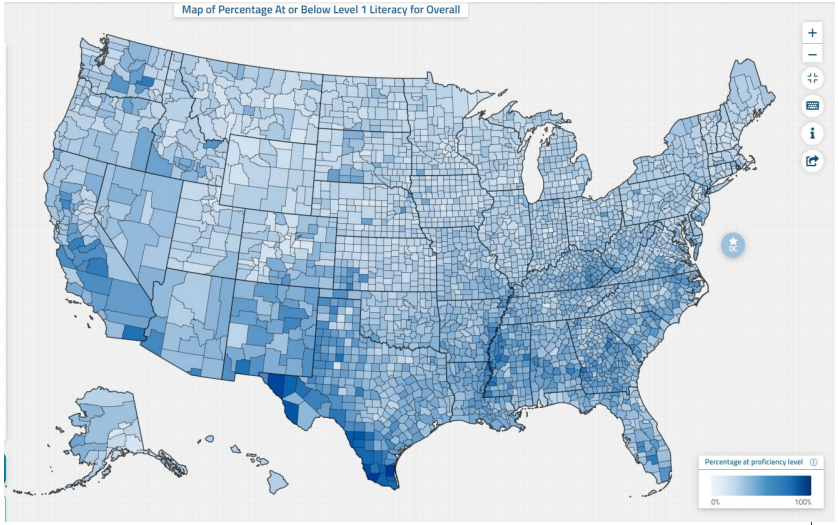
* Significantly different ($p < .05$) from the United States.

NOTE: LIT = Literacy. NUM = Numeracy. DPS = Digital problem solving. Average scores for the United States are compared to the PIAAC international average and highest-performing countries. Results for 23 of the countries were gathered in 2011–12, and an additional 9 participated in 2014–15. The two highest-performing countries are shown, in descending order from left to right within each domain. Results for the United States are shown on the far left within each domain to highlight that comparison.

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

https://nces.ed.gov/surveys/piaac/measure.asp?section=2&sub_section=5

Model-based state and county-level estimates Overall and by domain



https://nces.ed.gov/surveys/piaac/skillsmap/

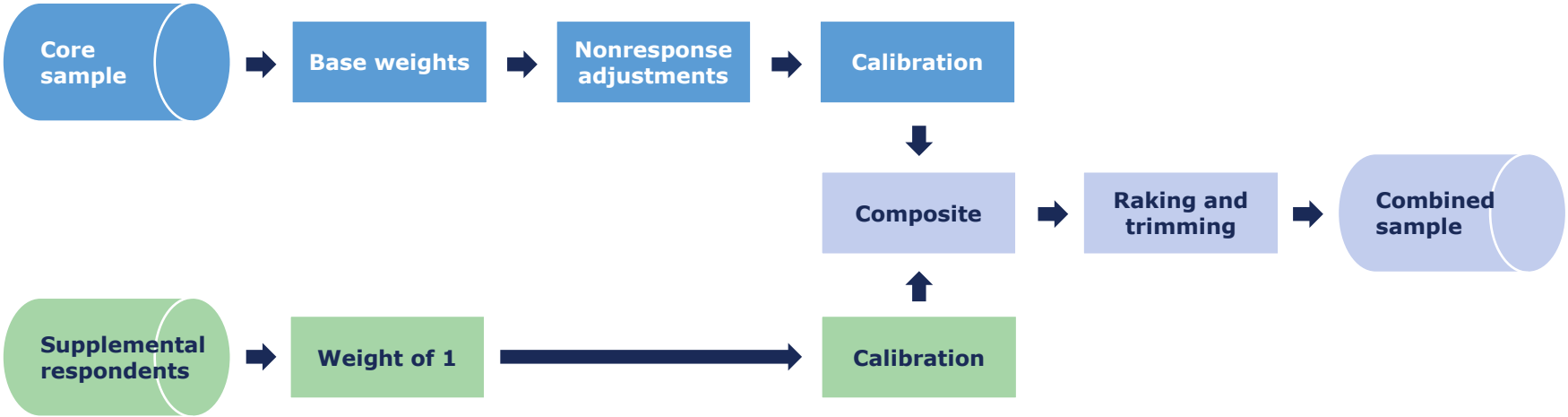
U.S. PIAAC Cycle II | Sample Design

- 4-stage area sample
- Two components
  - Core sample | nationally representative
  - State supplemental sample | additional PSUs (counties/groups of counties) so that combined sample has at least 2 PSUs per state
  - Intentionally separated in case we needed to stop the supplement
- Data collection for state supplement halted mid-field-period
  - 350 respondents (around 20% of target for supplemental sample)
  - Not evenly worked; work had not started in some PSUs
- Core sample | 4,287 respondents

Method for combining samples

Process

Calibration and compositing | based on Dever (2018)



Rationale

- Restrictions
  - Do not know the real probabilities due to stopped work
  - Need to produce weights (as opposed to estimates)
  - Do not have the outcome (proficiency scores) available at the time weighting

→ Model-based prediction (e.g., mass imputation) and doubly robust estimation are not applicable

- Small non-probability sample combined with large probability sample
  - Bias less a concern than with large non-probability samples
  - Placed emphasis on limiting variance

→ Base weight of 1 plus one-step calibration for supplemental sample

Pre-compositing calibration

- Post-stratified both samples to ACS population totals for age group by region by education
  - **Age** | To distinguish international target population vs U.S.-only
  - **Region** | Related to proportion of supplemental cases that were worked
  - **Education** | Related to proficiency

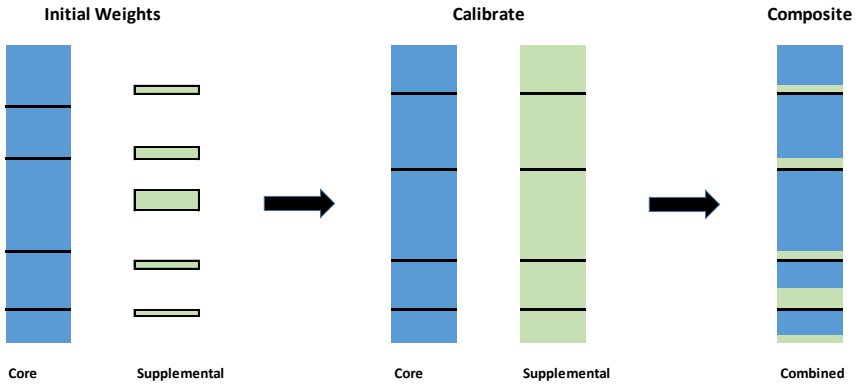
Cell	Age group	Region	Education level
1	16-65	1	-
2	16-65	2	-
3	16-65	3	Non-college graduates
4	16-65	3	College graduates
5	16-65	4	Non-college graduates
6	16-65	4	College graduates
7	66-74	-	Non-college graduates
8	66-74	-	College graduates

Compositing (1)

- Composite weights for person *i* in domain *g* (in non-certainty PSUs)

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

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



Compositing (2)

- Compositing factor
  - Based on Krenzke and Mohadjer (2020)

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

  - Attempts to give more weight to the sample with lower MSE
  - $n/(1+cv^2)$ = effective sample size → reflects variance
  - d = Kolmogorov-Smirnov statistic for detailed (9 category) educational attainment distribution (sample vs ACS) → reflects bias
  - g* = age group (16-65 or 66-74)

Compositing (3)

Domain (g)	n_g^C	n_g^S	$1+(cv_g^C)^2$	$1+(cv_g^S)^2$	d_{g*}^C	d_{g*}^S	α_g^C
1	463	29	1.30	1.00	0.017	0.075	0.928
2	554	34	1.27	1.00	0.017	0.075	0.931
3	1562	110	1.35	1.02	0.017	0.075	0.919
4	711	110	1.39	1.00	0.017	0.075	0.831
5	768	67	1.29	1.03	0.036	0.084	0.906

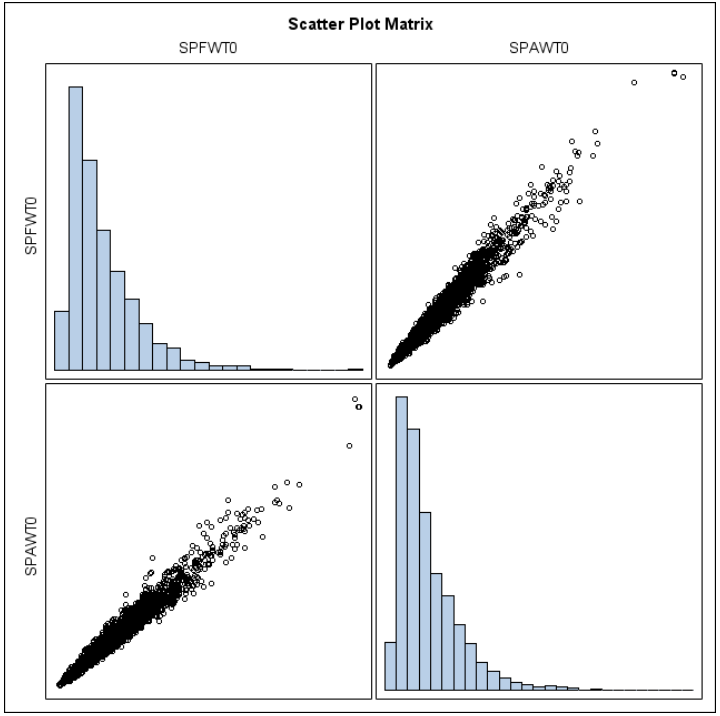
Evaluation/outcomes

National estimates | Evaluation (1)

- Evaluation was required to determine if the supplemental sample could be included in national and international reports
- Produced weights for core-only sample and compared results against composite sample in terms of:
  - Weights
  - Estimates
  - Variances
  - Associations

National estimates | Evaluation (2)

- **Weights** | correlation of 0.989

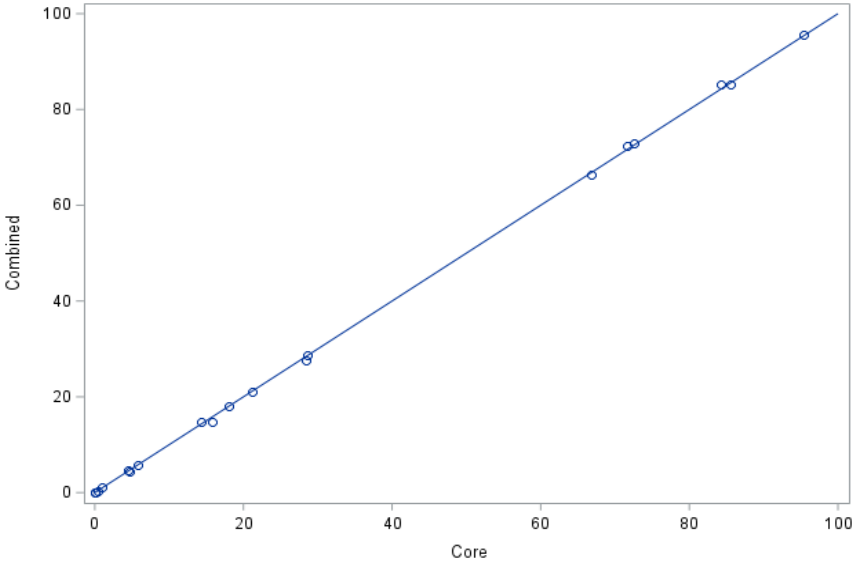


SPAWT0 = core-only weight; SPFWT0 = combined sample weight

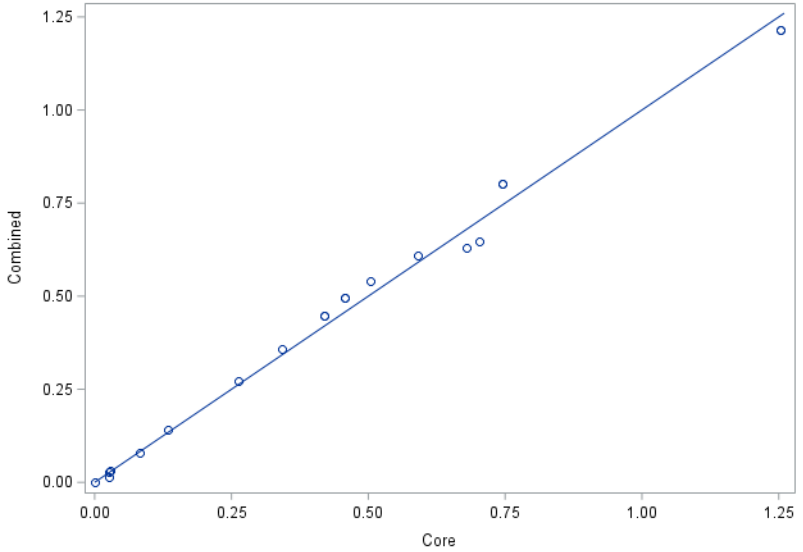
National estimates | Evaluation (3)

- **Estimates and variances** | weighted proportions for survey variables, e.g., computer experience, language spoken at home, financial literacy

Survey variable estimates for Core vs. Combined



Standard error estimates for Core vs. Combined



National estimates | Evaluation (4)

- **Associations**
  - Chi-square tests of associations between auxiliary variables
    - Among 21 tests, agreement on 20 as to whether or not association was statistically significant based on $\alpha = 0.05$
  - Regression - three auxiliary variables on education
    - R^2 was 11% for both models (core and composite)
    - Three auxiliary variables were statistically significant in both models

Models

- **IRT model** | combined sample met minimum sample size requirements
- **SAE** | Supplemental sample increased number of states and counties with data

→ Less reliance on the model

Count	Core	Combined
States with data	34	48
Counties with data	89	126
States with 2 or more PSUs	31	37

Summary and conclusion

Summary and conclusion

- **Challenge** | Incomplete supplemental sample
- **Method** | Calibration and compositing
- **Conclusion** | Including the incomplete supplemental sample in PIAAC
  - Has a negligible effect on national estimates
  - Strengthens model-based estimates

Thank you

WendyVandeKerckhove@westat.com

References

- Dever, J. A. (2018), "Combining Probability and Non-probability Samples to form Efficient Hybrid Estimates," Presentation Slides for the Federal Committee on Survey Methodology Conference, Washington, DC.
- Krenzke, T., and Mohadjer, L. (2020). Application of probability-based link-tracing and non-probability approaches to sampling out-of-school youth in developing countries. Journal of Survey Statistics and Methodology. doi: https://doi.org/10.1093/jssam/smaa010
- Van de Kerckhove, W., and Krenzke, T. (2022, August). Evaluating a sample design for small area estimation and adaptive field efforts (paper presenter). Joint Statistical Meetings, Washington, D.C.