An Examination of the Performance of Variance Estimators in ILSAs

Umut Atasever, Sabine Meinck, Diego Cortes

IEA General Assembly, Madrid, October 7, 2025

*This methodological research was possible thanks to your support of IEA's R&D fund

Content

- Background
- Motivation for this study
- RQ's
- Methods
- Results
- Discussion

Why this research matters?

TIMSS PIRLS ICCS ICILS

Because **IEA studies** guide education policy,
our results must be **accurate and precise** to remain credible

Why uncertainty matters?

Weather forecast 🌤️
20 °C tomorrow

19–21 °C vs. **10–30 °C**
→ *very different forecasts*

Why uncertainty matters?

Weather forecast 🌤️
20 °C tomorrow

19–21 °C vs. **10–30 °C**
→ *very different forecasts*

Math Achievement 📊
500

495–505 vs. **475–525**
→ *very different estimates*

Why uncertainty matters?

Weather forecast 🌤️
20 °C tomorrow

19–21 °C vs. **10–30 °C**
→ *very different forecasts*

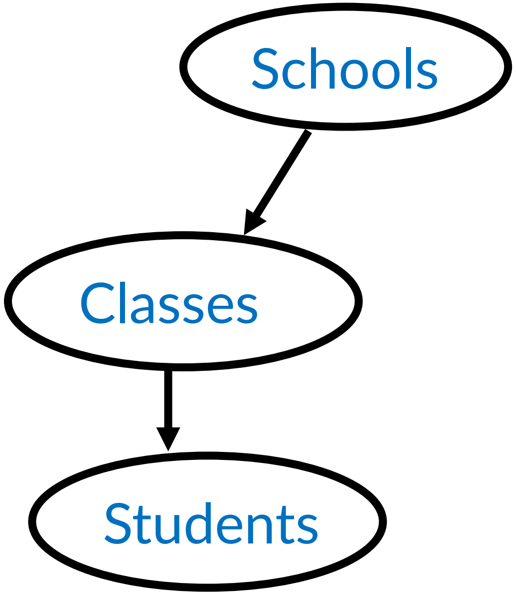
Math Achievement 📊
500

495–505 vs. **475–525**
→ *very different estimates*

Uncertainty in estimates arises from **both assessment and sampling design** in IEA studies

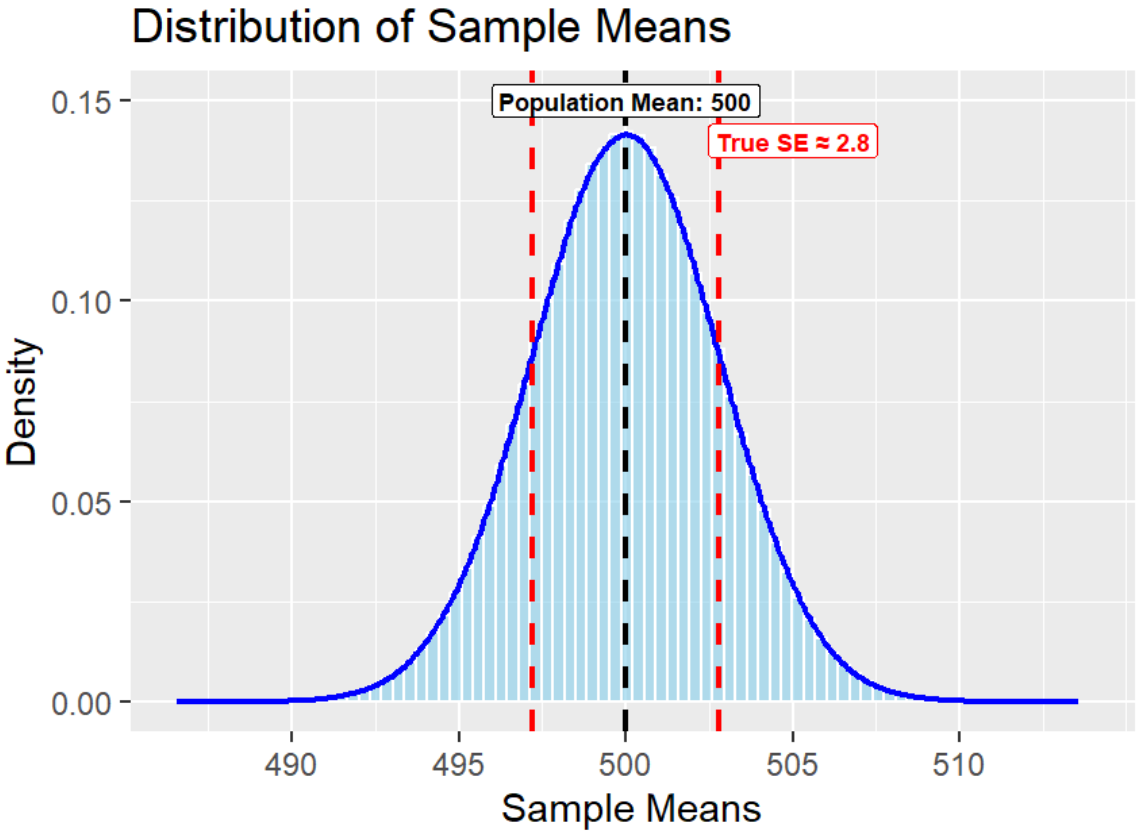
- Today, we focus on **sampling design**

Sampling Design in IEA Studies



- We carefully **sample schools and classes**
- We then assess **students** in the sampled classes
- Not all students or schools take part (**non-response**)
- **The final sample** is only one of many possible samples
- **Uncertainty** from these processes leads to **sampling error**

Focusing on the sampling error (SE)



- **SE** = the variation of results across all possible samples
- If we had all possible samples, we could compute **the true SE**
- With **only one sample**, we must estimate **the SE**
  - That is where **replication methods** come in

Replication Methods

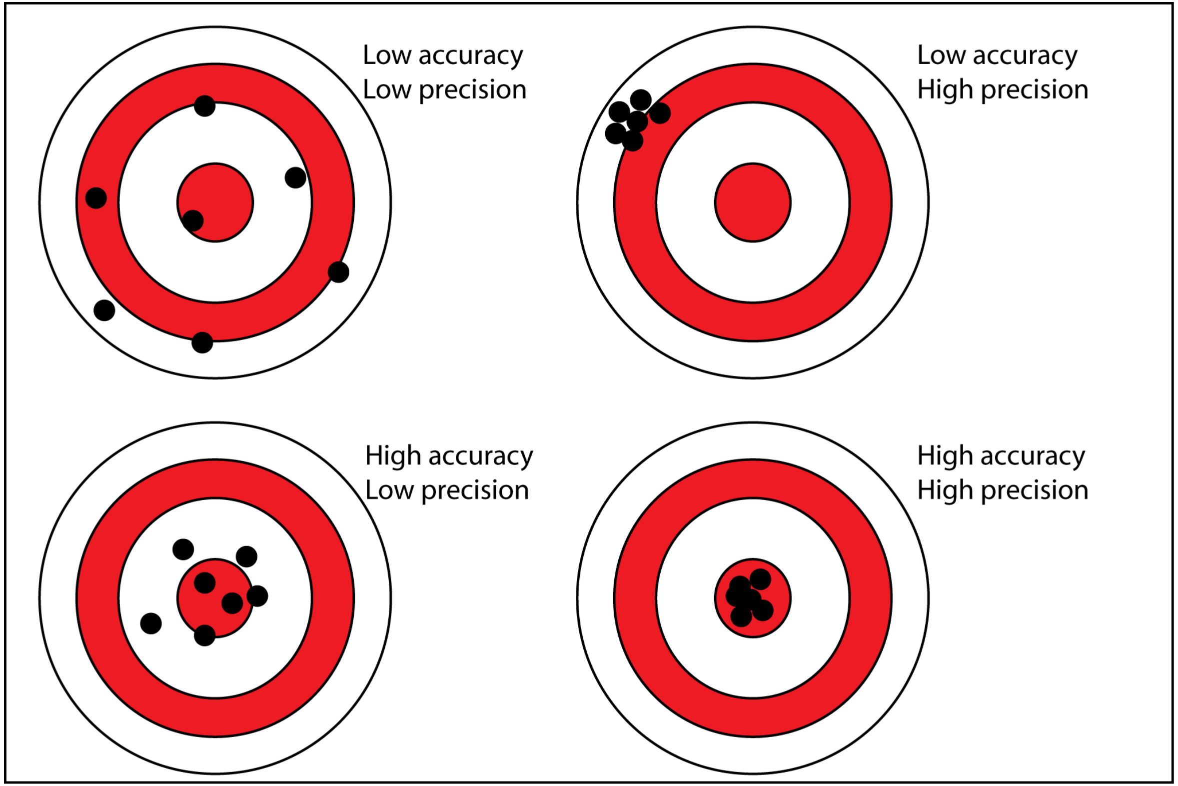
Replication ≈ resampling to approximate the sampling distribution

- Creates many sub-samples (replicates) from one sample
- Handles complex designs

Main approaches:

- **Jackknife Repeated Replication (JK2)**
  - ➢Used in IEA studies with
    - ▪ **Half** variant (e.g., 125 replicates)
    - ▪ **Full** variant (e.g., 250 replicates)
- **Balanced Repeated Replication (BRR)**
  - ➢Used in OECD studies, often with a Fay adjustment (e.g., 0.5)
- **Bootstrapping**
  - ➢Common in social sciences

Replication Methods – how do we compare them?



Accuracy → closeness to the true sampling error

Precision → consistency across samples

Gap in Literature

- No systematic comparisons of the replication methods in ILSAs
- Most evidence comes from other fields (e.g., Valiant et al., 2018)

Gap in Literature and Research Questions

- No systematic comparisons of the replication methods in ILSAs
- Most evidence comes from other fields (e.g., Valiant et al., 2018)

Which replication method works best for…

- Smooth statistics (means, correlations)
- Non-smooth statistics (percentiles, benchmarks)

How do methods perform under different conditions?

- Population backgrounds
- Sample sizes
- Non-response

Methodology

- 👥 **Population Design:** simulated a population with TIMSS-like characteristics
- 🔎 **Sampling Design:** two-stage stratified cluster sampling (schools and classes)
- 🖥️ **Simulation Procedure:** 1,000 samples per scenario (sampling distribution)
- 🧮 **Replication Methods (13):** JK2 (8), BRR (4), Bootstrapping (1)
- 📈 **Statistics**
  - **Smooth:** Means, subgroup differences, correlations
  - **Non-smooth:** Percentiles, benchmarks
- 🔢 **Total Iterations** about **1.56 million**

IEA

Results – Highlights (1)

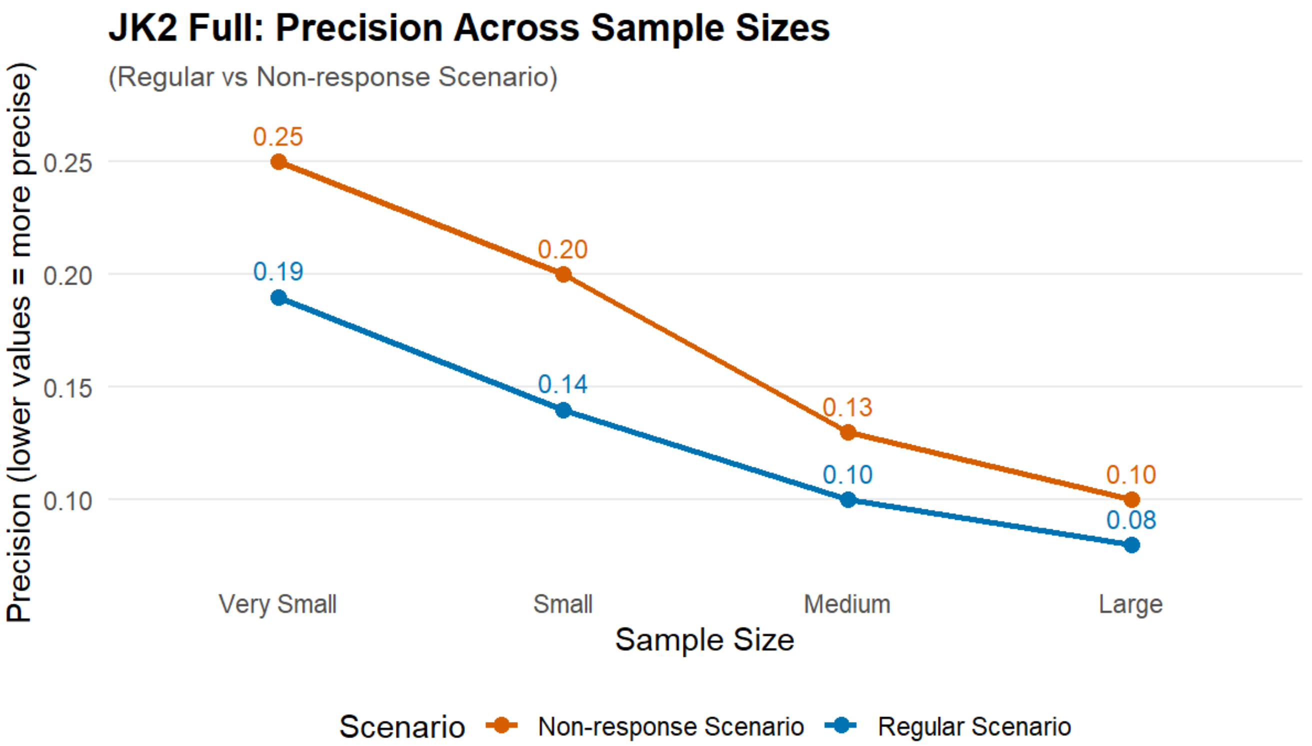
- **Methods perform well** under typical IEA sampling designs, including **JK2**, which we use in IEA studies
- **Smooth statistics** (means, correlations):
  - ➢**JK2 (Half or Full)** and **BRR** are both accurate and stable
- **Non-smooth statistics** (percentiles):
  - ➢**BRR** generally outperforms **JK2**
  - ➢**BRR with lower Fay factors (0.1)** performs much better than the **0.5 Fay** used in OECD studies
- **JK2 Half** performs same as **JK2 Full**
  → but requires less computation

Results – Highlights (2)

- **Non-response** reduces precision across all methods, especially in small samples
  → This is a real-world challenge and shows why **high participation is critical**

- **Re-pairing participating schools after non-response** improves accuracy
  → **IEA** uses this approach, while **PISA** keeps original order, which risks bias

Precision Under Non-response



Discussion – Key Takeaways

✅ **Practical guidelines** now exist to help IEA choose the most suitable replication methods

✅ **IEA replication methods are robust** — they deliver accurate and precise results under typical study conditions

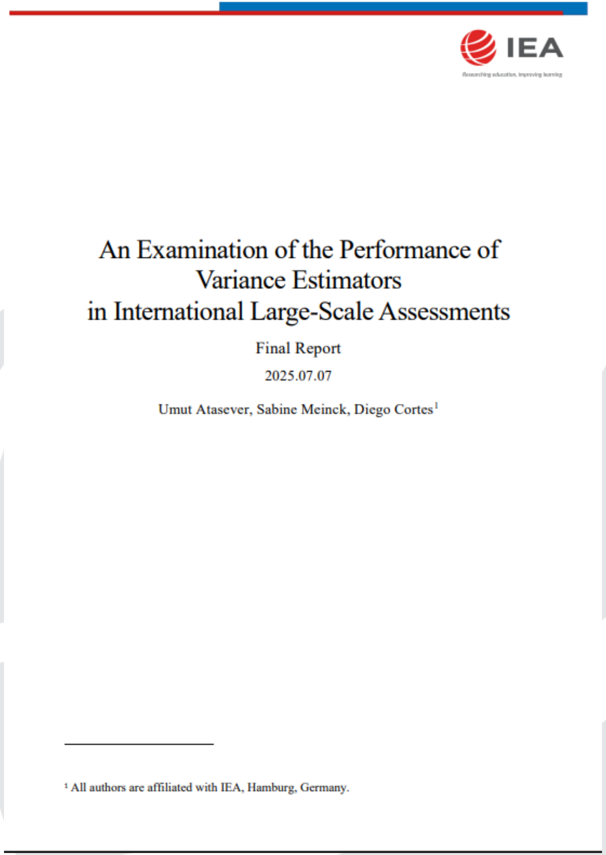
🚨 **Non-response is the biggest challenge** – reducing non-response is key to protecting accuracy and precision

👉 **Implication:**

- Method choice should depend on **statistic type** and **sample conditions**
- No one-size-fits-all, but **IEA approaches** have proven **reliable** across scenarios

The comprehensive report, with all scenarios and detailed results, can be found on the IEA website

QR Code


IEA

Researching education, improving learning

An Examination of the Performance of Variance Estimators in International Large-Scale Assessments

Final Report

2025.07.07

Umut Atasever, Sabine Meinck, Diego Cortes[1]

[1] All authors are affiliated with IEA, Hamburg, Germany.

Contact: umut.atasever@iea-hamburg.de

References

Beaumont, J.-F., & Émond, N. (2022). A Bootstrap Variance Estimation Method for Multistage Sampling and Two-Phase Sampling When Poisson Sampling Is Used at the Second Phase. *Stats*, 5(2), 339-357. https://doi.org/10.3390/stats5020019

Bruch, C., Munnich, R., & Zins, S. (2011). Variance Estimation for Complex Surveys. *Advanced Methodology for European Laeken Indicators.*

Dippo, C. S., Fay, R. E., & Morganstein, D. H. (1984). Computing variances from complex samples with replicate weights. In *Proceedings of the Survey Research Methods Section* (pp. 489-494). Alexandria, VA: American Statistical Association.

Efron, B. (with Society for Industrial and Applied Mathematics). (1982). The jackknife, the *Bootstrapping*, and other resampling plans. *Society for Industrial and Applied Mathematics* (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104). https://doi.org/10.1137/1.9781611970319

IEA. (2022). Help Manual for the IEA IDB Analyzer (Version 5.0). Hamburg, Germany. (Available from www.iea.nl)

Judkins, D. (1990). Fay's Method for Variance Estimation. Statistics Sweden, *Journal of Statistics Sweden*, Vol 6., No.3, 223–239.

Kovar, J. G. (1985). Variance Estimation of Nonlinear Statistics in Stratified Samples. Methodology Branch Working Paper No. 85-052E, Statistics Canada.

Kovar, J. G., Rao, J. N. K., & Wu, C. F. J. (1988). Bootstrapping and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16(S1), 25–45. https://doi.org/10.2307/3315214

Krewski, D., & Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. T*he Annals of Statistics*, 9(5), 1010–1019. https://doi.org/10.1214/aos/1176345580

Mang, J., Küchenhoff, H., Meinck, S. (2024). Evaluating German PISA stratification designs: a simulation study. Large-scale Assessments in Education (2024) 12:15, https://rdcu.be/dHn8h

Rust, K., & Rao, J. (1996). Variance estimation for complex surveys using replication techniques. Statistical *Methods in Medical Research*, 5(3), 283–310. https://doi.org/10.1177/096228029600500305

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment* (pp. 117–130). Taylor & Francis. https://doi.org/10.1201/b16061-11

Valliant, R., Dever, J. A., & Kreuter, F. (2018). Practical Tools for Designing and Weighting Survey Samples. *Springer International Publishing*. https://doi.org/10.1007/978-3-319-93632-1

Wolter, K. M. (2007). Introduction. In Introduction to Variance Estimation (pp. 1–20). *Springer New York*. https://doi.org/10.1007/978-0-387-35099-8_1

Woodruff, R. S. (1952). Confidence Intervals for Medians and Other Position Measures. *Journal of the American Statistical Association*, 47(260), 635–646. https://doi.org/10.1080/01621459.1952.10483443