

Utilizing R for Simulating Studies in International Large-Scale Assessments in Education

Umut Atasever (IEA, Hamburg, Germany)

Francis L. Huang (University of Missouri)

Leslie Rutkowski (Indiana University)



International Large-Scale Assessments (ILSAs)

- IEA (International Association for the Evaluation of Educational Achievement) conducts large-scale, comparative studies in the field of education
- Goal: Gain a deep understanding of the effects of educational policies and practices on student achievement.
- ILSA measure: Student achievement in subjects such as math, reading, and civic education.
- Background information about students (e.g., attitudes, home support), schools (e.g., resources, instructional practices), and teachers.

TIMSS (Trends in International Mathematics and Science Study)

- Upcoming Release: TIMSS 2023 – December 4, 2023
- Measures trends in student achievement in mathematics and science across countries in grade 4 and grade 8
- 72 countries, 28 years of trend data with 8 data points

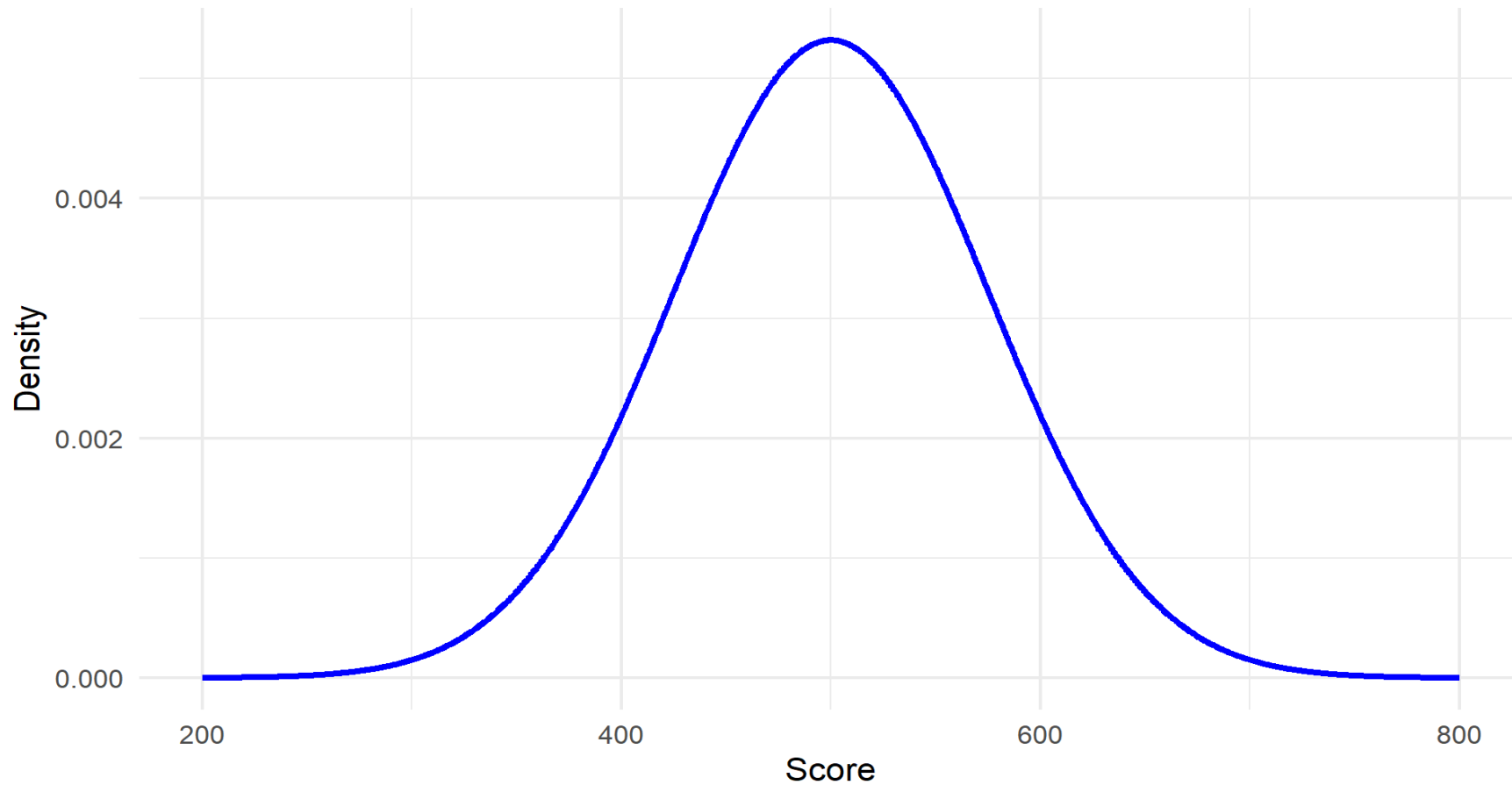
ICILS (International Computer and Information Literacy Study)

- Recent Release: ICILS 2023 – November 12, 2023
- Measures student proficiency in computer and information literacy to understand how well students are prepared for the digital world in grade 8
- 35 countries, 10 years of trend data with 3 data points









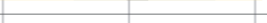











Visit IEA's website for further studies: <https://www.iea.nl/studies>

R data is also available for these studies

Normal Distribution (Mean = 500, SD = 75)

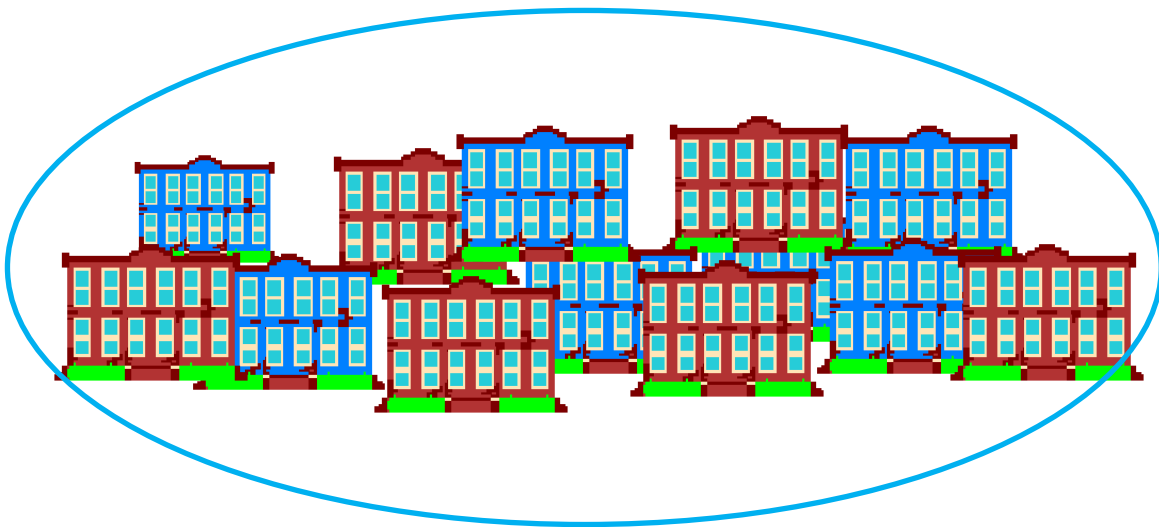


Computer Literacy Scores in ICILS 2023

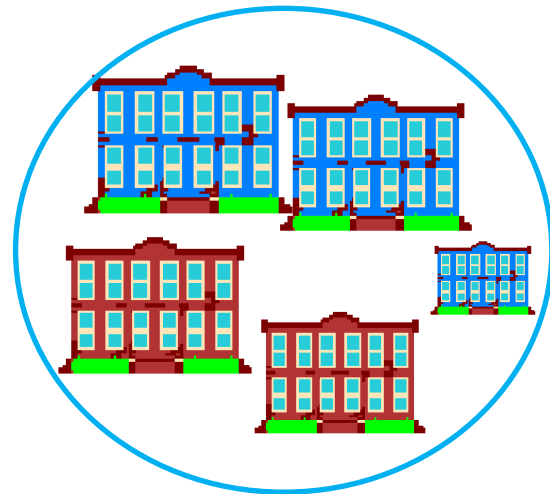
Country	Average CIL scale score					CIL distribution
† Korea, Republic of	540 (2.5)	▲				
¹ Czech Republic	525 (2.1)	▲				
† ¹ Denmark	518 (2.7)	▲				
Chinese Taipei	515 (3.0)	▲				
† Belgium (Flemish)	511 (4.4)	▲				
¹ Portugal	510 (3.0)	▲				
¹ Latvia	509 (3.6)	▲				
Finland	507 (3.6)	▲				
¹ Austria	506 (2.5)	▲				
Hungary	505 (3.8)	▲				
¹ Sweden	504 (3.0)	▲				
¹ Norway (Grade 9)	502 (2.9)	▲				
Germany	502 (3.5)	▲				
Slovak Republic	499 (2.7)	▲				
France	498 (2.7)	▲				
¹ Spain	495 (1.9)	▲				
Luxembourg	494 (2.0)	▲				
Italy	491 (2.6)	▲				
¹ Croatia	487 (3.9)	▲				
¹ Slovenia	483 (2.3)	▲				

Two-stage design – Stage 1

Within each explicit stratum, schools are sampled with probabilities proportional to their size; 150 schools

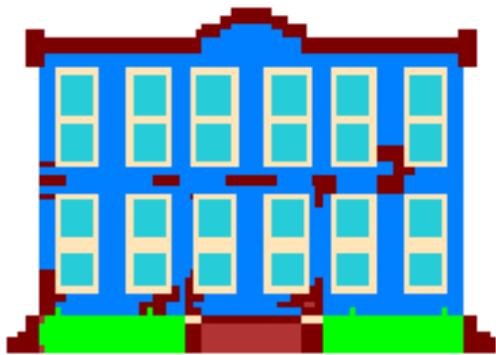


Population



Sample

Two-stage design – Stage 2



Population of
students / classes
within participating
schools



Sample of students /
classes within
participating schools
e.g, 3000-5000
students

Why are ILSAs complex?

- Different sampling strategies impact the complexity of the data
 - Probabilistic two-stage design
 - Sample size
 - Stratification
 - Oversampling (e.g., private schools)
 - Nonresponse adjustments due to school and student unit non-participation
- There are multi-stage weights that make analyses more complex (i.e., school weights, student weights, total final weights)

The need for simulation studies

- Analyzing the data becomes tricky, especially when using two-level models (i.e., multilevel models)
- Studies cannot determine which method is better without true population values
- Simulations allow us to compare results with true population values
- Real-world ILSAs have complexities like nonresponse and different sampling methods

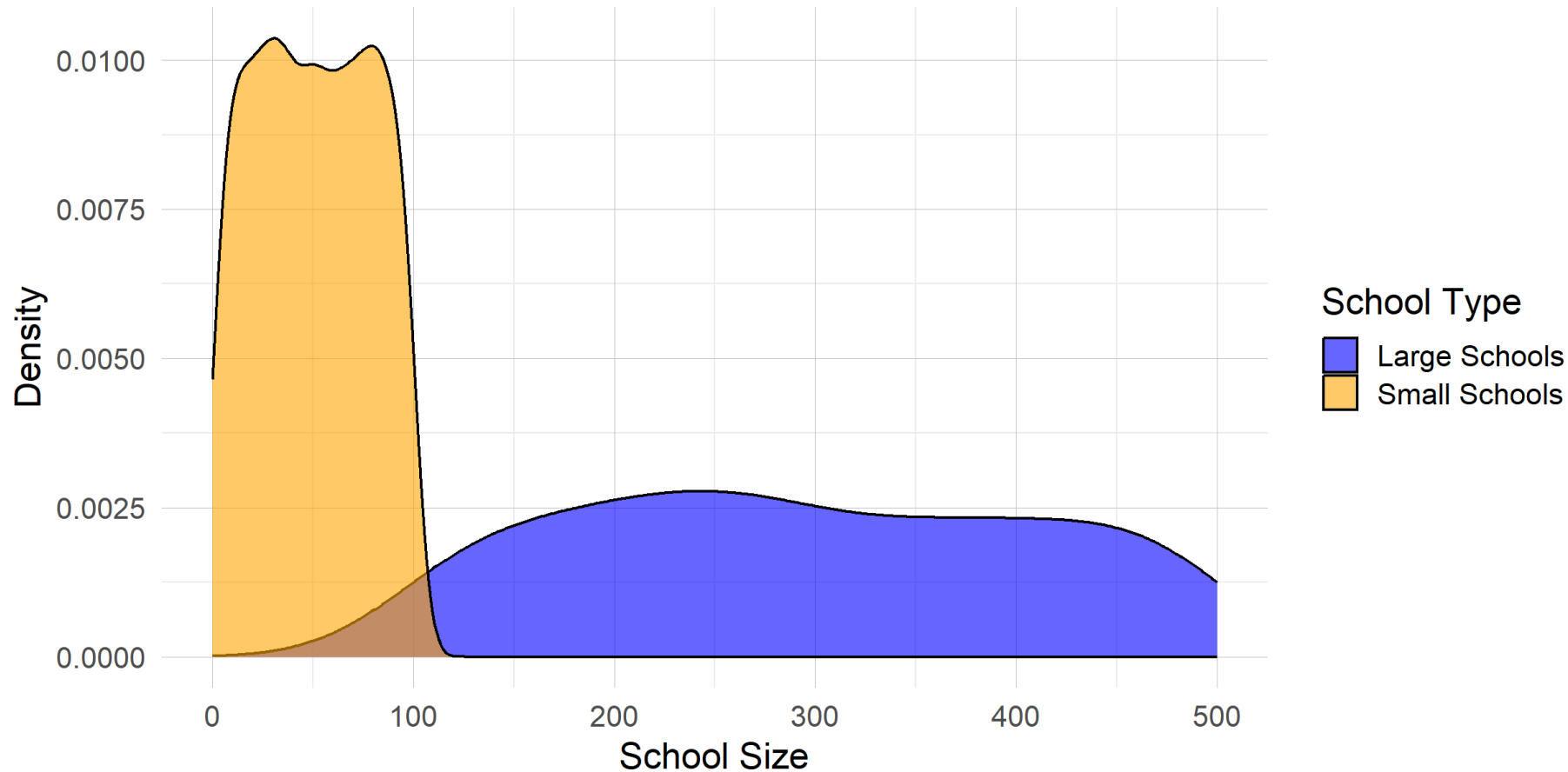
Objective

- **Goal:**
 - Introduce a simulation approach using R
 - Make simulations understandable and easy to replicate
 - Be efficient with multicore computation
- **Method:**
 - Generate a synthetic finite population with known parameters
 - Simulate sampling designs similar to ILSAs (1000 times)
 - Clustered data – the Intra Class Coefficient (ICC) typically between 25% and 50%
 - This design can be also applicable other fields

Simulation Framework

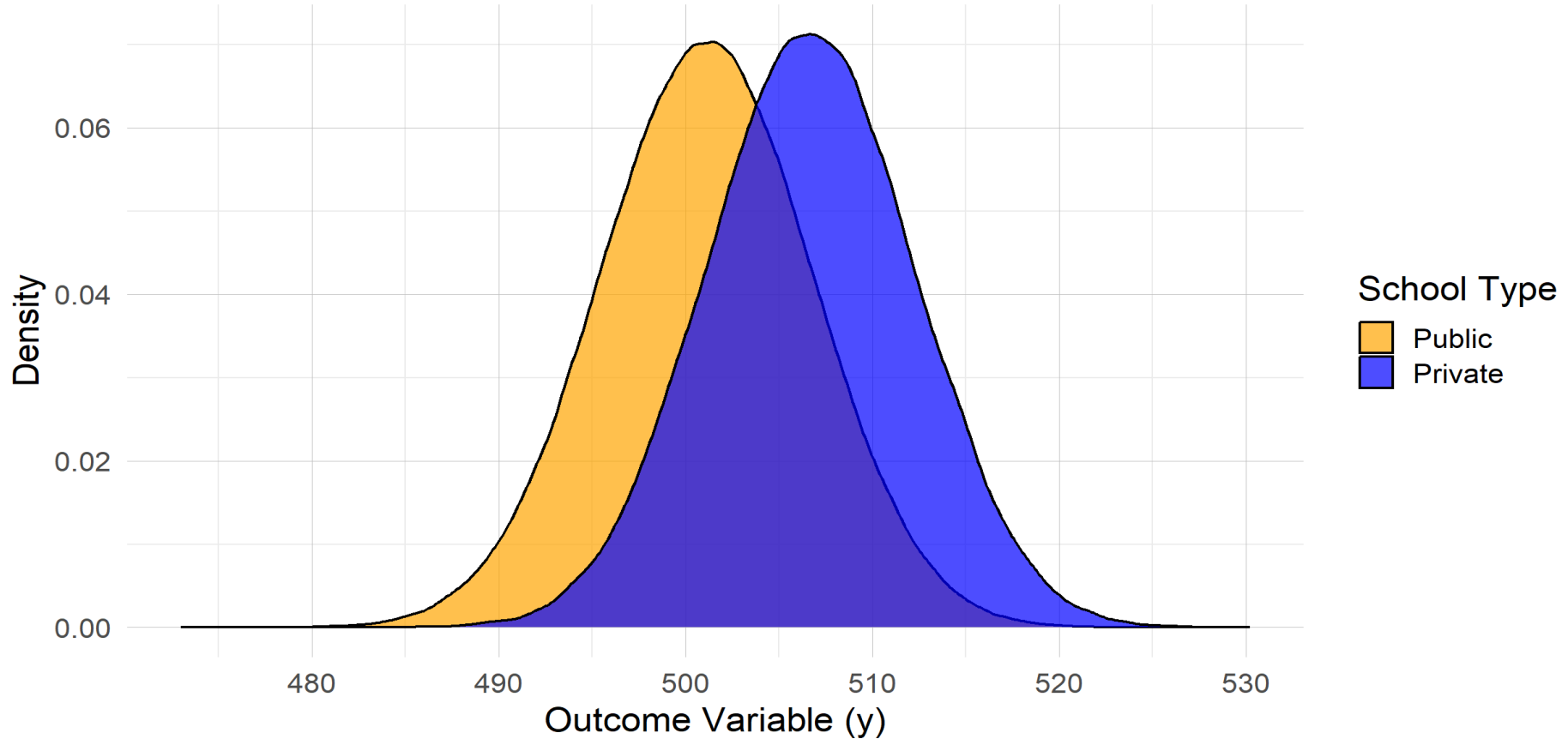
- Generate finite population dataset (10,000 schools, ~972,000 students)
- Apply stratified two-stage sampling procedures for 1000 times
 - Add further complexities (e.g., stratification, non-response)
- Analyze 1,000 samples using different weighting methods
- Fit mixed models with different weight configurations
 - Level 1 (students) only
 - Level 2 (schools) only
 - Two-level weights
 - No weights

Density Distribution of School Sizes (Large vs Small Schools)



Distribution of Outcome Variable (y)

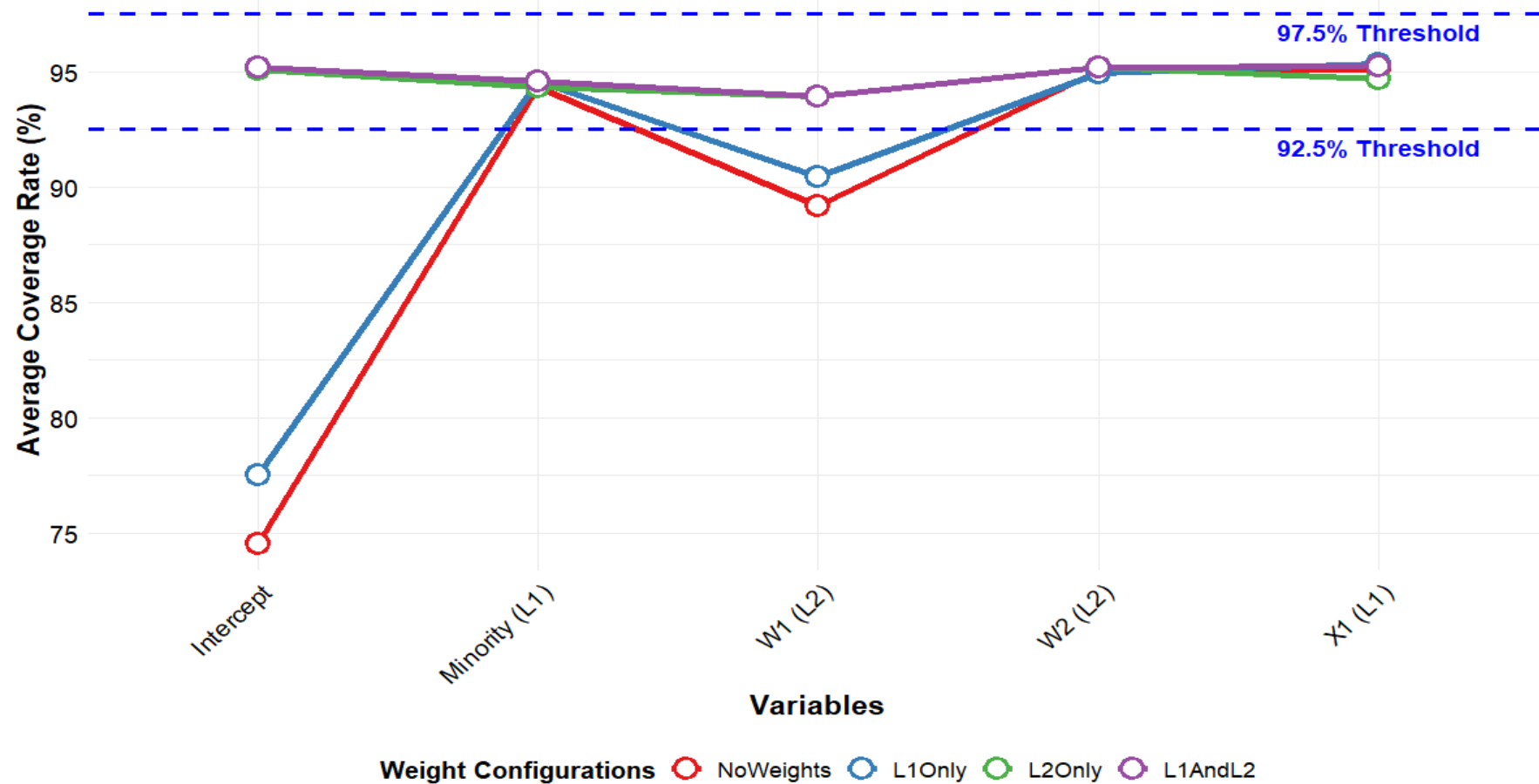
$$y = 500 + w_1 \times 3 + \text{private} \times 3 + w_2 \times 3 + x_1 \times 3 + \text{minority} \times -3 + e_2 + e_1$$



Coverage rates

- Proportion of confidence intervals that cover the true population value.
- **Ideal Coverage:** Approximately 95% (Bradley's 1978 liberal criterion)
- **Acceptable Thresholds:** 92.5% – 97.5% coverage
- **Why it Matters:**
- **Too High Coverage:** Type II errors (missing true effects)
- **Too Low Coverage:** Type I errors (false positive results)

Average Coverage Rates across Weight Configurations and Variables



Summary

- The methods discussed, especially simulation-based techniques, can be applied to compare different sampling strategies, weights, and variance estimation methods
- If you have a methodological work that can be applied in this field:
- You are welcome to submit your work to our open access journal:

<https://largescaleassessmentsineducation.springeropen.com>

- and, to attend the IEA International Research Conference, in Rome from June 25–27, 2025.

Thank you for your attention!

- For further comments or questions, please contact me:

umut.atasever@iea-hamburg.de

R Code

Conditions tested

Condition	Levels	Description
Number of Clusters (J)	100, 150, 200	Variations in the number of clusters (schools) sampled
School Nonresponse (SCH_NR)	TRUE, FALSE	Indicates whether school-level nonresponse is present
Student Nonresponse (ST_NR)	TRUE, FALSE	Indicates whether student-level nonresponse is present
Classroom Sampling (CS)	TRUE, FALSE	Indicates whether classroom sampling is used
Population ICC	0.50 (Large), 0.25 (Moderate)	Varying Intraclass Correlation Coefficient (ICC), representing between-group variance
Replications	1,000	Number of replications used to test the conditions for each setup

Generating data – Step 1

```
# Number of schools
j2 <- 1000 # Large schools
j1 <- 9000 # Small schools

totj <- j2 + j1

# Total number of schools
# Create a data frame for schools dfr <- data.frame(id = 1:totj)

# Set random seed for reproducibility
set.seed(1112)

# Simulate school sizes
dfr$size[1:j2] <- sample(100:500, size = j2, replace = TRUE) # Large schools
dfr$size[(j2 + 1):totj] <- sample(1:100, size = j1, replace = TRUE) # Small schools
```

Generating data – Step 2

```
# Simulate school-level variables
dfr$w1 <- rnorm(totj, 0, 1) # Continuous predictor (w1)
dfr$w2 <- rbinom(totj, 1, plogis(scale(dfr$size) - .25)) # Binary predictor based on
size
dfr$private <- rbinom(totj, 1, plogis(dfr$w1 - 2.5)) # Private school indicator
(based on w1)
dfr$e2 <- rnorm(totj, 0, 2) # School-level error term (e2)

# Expand school-level data to student-level data

dat <- data.frame( id = rep(dfr$id, dfr$size), size = rep(dfr$size, dfr$size), private
= rep(dfr$private, dfr$size), w1 = rep(dfr$w1, dfr$size), w2 = rep(dfr$w2,
dfr$size), e2 = rep(dfr$e2, dfr$size))
```

Generating data – Step 3

```
# Define level-1 error term (e1) to control ICC (Intraclass Correlation)

sigm <- 2.77 # For ICC = 0.25
dat$x1 <- rnorm(Ns) # Independent variable x1
dat$minority <- rbinom(Ns, 1, 0.1) # Binary minority variable (prevalence ~10%)
dat$e1 <- rnorm(Ns, 0, sigm) # Student-level error term

# Combine the predictors to make the outcome variable 'y'
dat$y <- 500 + dat$w1 * 3 + dat$private * 3 + dat$w2 * 3 + dat$x1 * 3 +
dat$minority * -3 + dat$e2 + dat$e1 # Add effects from all predictors and errors
```

Generating data – Step 4

```
# Fit a random intercept model (null model)

library(lme4)

m0 <- lmer(y ~ (1 | id), data = dat)

# Summary of the model and ICC calculation

summary(m0) # Model summary

performance::icc(m0) # ICC (Intraclass Correlation) calculation
```

```
# Intraclass Correlation Coefficient Adjusted ICC: 0.249 Unadjusted ICC: 0.249
```

Sampling from this population

```
# Stratified PPS sampling for 150 schools  
  
sampled_schools <- rbind(  
  
  public_schools[sample(1:nrow(public_schools), 120, prob = public_schools$size), ],  
  private_schools[sample(1:nrow(private_schools), 30, prob = private_schools$size), ]  
)
```

- Repeat it for 1000 times further with different sampling scenarios

Model specifications

Main model with varying weights

```
model <- mix( y ~ w1 + w2 + x1 + minority + (1 | id), data = sampled_data, weights  
= c('x', 'y'), cWeights = TRUE )
```

No weights at both levels

```
weights_no = c('one', 'one')
```

School weights only (level 2)

```
weights_school = c('one', 'schwgt')
```

Level 1 weights only (normalized)

```
weights = c('nwt', 'one')
```

Both levels weights

```
weights = c('stdwgt', 'schwgt')
```

Parallel processing with library (parallel)

```
# Check available cores and create cluster
num_cores <- detectCores() - 1
cl <- makeCluster(num_cores)
# Define variables on each worker
clusterEvalQ(cl, {n_sample <- 150; dir <- "Z"})
# Export scripts and run in parallel
clusterExport(cl, c("run_script", "script1", "script2", "script3"))
parLapply(cl, list(script1, script2, script3), run_script)
# Stop the cluster
stopCluster(cl)
```