Oxford University Centre for Educational Assessment

UNIVERSITY OF OXFORD

Complex designs of international assessments

Professional Development and Training Course:
Analyzing International Large-Scale Assessment Data with R

Dr. Daniel Caro & Dr. Christian Bokhove

AERA 2014
Philadelphia, April 2, 2014

UNIVERSITY OF Southampton

Outline

- International assessment studies
- Two-stage sampling
  - Replicate weights
- Rotated test design
  - Plausible values
- Handling complex designs

Main studies

- Three largest studies:
  - PISA (Programme for International Student Assessment)
  - TIMSS (Trends in International Mathematics and Science Study)
  - PIRLS (Progress in International Reading Literacy Study)
- Organisations involved
  - IEA (TIMSS and PIRLS)
  - OECD (PISA)

Data and instruments

- Studies collect information about students, their families, teachers, and schools in a number of educational systems.
- Instruments
  - Achievement test
  - Student questionnaire
  - Parent questionnaire
  - Teacher questionnaire
  - School questionnaire

Overview of studies

	PISA	TIMSS	PIRLS
Subjects tested	Reading, mathematics, science	Mathematics, science	Reading
Target population	15-year-olds	Grade 4 and 8	Grade 4
Available cycles	5, from 2000 to 2012	5, from 1995 to 2011	3, from 2001 to 2011
Frequency	Every 3 years (emphasis on one subject in each cycle)	Every 4 years	Every 5 years

Where to find the data

- PISA
  - http://pisa2012.acer.edu.au/downloads.php
  - SAS, SPSS control files and txt format
- TIMSS and PIRLS
  - http://rms.iea-dpc.org/
  - SAS and SPSS data files
  - Codebooks, reports, user guides

Complex study designs

- Two-stage sampling
  - 1 stage: school selection
  - 2 stage: student selection
- Rotated test design
  - Booklets do not include all items to reduce testing time
  - Linking items make it possible to assign scores to students as if they had responded to all items

Outline

- International assessment studies
- Two-stage sampling
  - Replicate weights
- Rotated test design
  - Plausible values
- Handling complex designs

Sampling in educational studies

- Random sampling is rarely used in educational surveys:
  - Too expensive (e.g., training test administrators and travel costs)
    - Selected students attend many different schools
  - It is not practical to contact many schools
  - A link with class, teacher, school variables is sought
- Sampling is usually conducted in two stages
  - Two-stage sampling

Two stages

- First stage
  - Schools are selected
- Second stage
  - Students (PISA) or classes (TIMSS/PIRLS) are selected
    - 35 students selected randomly (PISA)
    - One or two intact classes (TIMSS/PIRLS)

Dependency of observations

- Students within schools tend to be more similar than across schools
  - Family background
  - Instructional setting
- Observations within schools are not independent
  - Dependency of observations yields less information about the population
  - Uncertainty is greater in two-stage sampling
- Assume two sampling strategies
  - 1000 students selected randomly
  - 10 schools selected randomly with 100 students each

Two extreme cases

- Case 1: students in the population are randomly assigned to schools
  - No differences between schools
  - Selecting 10 schools with 100 each is similar to selecting 1000 students randomly
- Case 2: students within schools are perfectly identical
  - 100 students in each school equivalent to 1 student
  - Sample of 10 students vs sample of 1000 students
- Dependency in real data lies between these two scenarios
  - The sample of students covers better the diversity of the population
  - Uncertainty (standard error) related with any parameter estimate is greater for two-stage sampling

Replicate weights

- Replicate weights or resampling techniques are used to calculate correct standard errors in two-stage sampling designs
- The idea behind:
  - There are many possible samples of schools and not all of them yield the same estimates
  - Use different samples of schools to calculate estimates
  - Take into account error of selecting one school and not another (sampling error)
- Each replicate weight represents one sample
- Variability between estimates reflects the sampling error

Two replication methods

- Jackknife
  - TIMSS and PIRLS
- Balanced repeated replication (BRR)
  - PISA uses a variant of BRR

Jackknife

- Schools are paired with other similar schools within zones
- A replicate is created for each zone or pair of schools
- One school is randomly removed within each zone and the weight of the other school is doubled
- Replicate estimates are compared with estimates for the total weight

Jackknife

Table 4.11

The Jackknife replicates for stratified two-stage sample designs

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	1	2	1	1	1	1	1	1	1	1	1
1	2	0	1	1	1	1	1	1	1	1	1
2	3	1	0	1	1	1	1	1	1	1	1
2	4	1	2	1	1	1	1	1	1	1	1
3	5	1	1	2	1	1	1	1	1	1	1
3	6	1	1	0	1	1	1	1	1	1	1
4	7	1	1	1	0	1	1	1	1	1	1
4	8	1	1	1	2	1	1	1	1	1	1
5	9	1	1	1	1	2	1	1	1	1	1
5	10	1	1	1	1	0	1	1	1	1	1
6	11	1	1	1	1	1	2	1	1	1	1
6	12	1	1	1	1	1	0	1	1	1	1
7	13	1	1	1	1	1	1	0	1	1	1
7	14	1	1	1	1	1	1	2	1	1	1
8	15	1	1	1	1	1	1	1	0	1	1
8	16	1	1	1	1	1	1	1	2	1	1
9	17	1	1	1	1	1	1	1	1	0	1
9	18	1	1	1	1	1	1	1	1	2	1
10	19	1	1	1	1	1	1	1	1	1	2
10	20	1	1	1	1	1	1	1	1	1	0

Source: OECD (2009). *PISA Data Analysis Manual: SPSS (2nd Edition. Paris)*: OECD Publishing.

Balanced repeated replicates

- Select one school at random within each stratum
- Set its weight to 0
- Double the weight of the other school

Balanced repeated replicates

Table 4.12

Replicates with the Balanced Repeated Replication method

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	1	2	0	0	2	0	0	0	2	2	2	0	2
1	2	0	2	2	0	2	2	2	0	0	0	2	0
2	3	2	2	0	0	2	0	0	0	2	2	2	0
2	4	0	0	2	2	0	2	2	2	0	0	0	2
3	5	2	0	2	0	0	2	0	0	0	2	2	2
3	6	0	2	0	2	2	0	2	2	2	0	0	0
4	7	2	2	0	2	0	0	2	0	0	0	2	2
4	8	0	0	2	0	2	2	0	2	2	2	0	0
5	9	2	2	2	0	2	0	0	2	0	0	0	2
5	10	0	0	0	2	0	2	2	0	2	2	2	0
6	11	2	2	2	2	0	2	0	0	2	0	0	0
6	12	0	0	0	0	2	0	2	2	0	2	2	2
7	13	2	0	2	2	2	0	2	0	0	2	0	0
7	14	0	2	0	0	0	2	0	2	2	0	2	2
8	15	2	0	0	2	2	2	0	2	0	0	2	0
8	16	0	2	2	0	0	0	2	0	2	2	0	2
9	17	2	0	0	0	2	2	2	0	2	0	0	2
9	18	0	2	2	2	0	0	0	2	0	2	2	0
10	19	2	2	0	0	0	2	2	2	0	2	0	0
10	20	0	0	2	2	2	0	0	0	2	0	2	2

Source: OECD (2009). *PISA Data Analysis Manual: SPSS (2nd Edition. Paris)*: OECD Publishing.

Balanced repeated replicates

- Each replicate sample only uses half of the available observations
- Sample size reduction could be problematic for subpopulations
- The Fay variant of BRR is applied
  - Instead of 2, weights are multiplied by 1.5
  - And instead of 0, weights are multiplied by 0.5

Balanced repeated replicates

Table 4.13

The Fay replicates

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	1	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5
1	2	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5
2	3	1.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5
2	4	0.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5
3	5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5
3	6	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5
4	7	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5
4	8	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5
5	9	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5
5	10	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5
6	11	1.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5
6	12	0.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5
7	13	1.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5
7	14	0.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5
8	15	1.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5
8	16	0.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5
9	17	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5
9	18	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5
10	19	1.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5
10	20	0.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5

Source: OECD (2009). *PISA Data Analysis Manual: SPSS (2nd Edition. Paris)*: OECD Publishing.

Sampling error

- Replicate weights account for the sampling error
- Estimates are calculated using each of the replicate weights
- The sampling error reflects the variability between estimates from different samples
- The formula can be applied to any statistic

Outline

- International assessment studies
- Two-stage sampling
  - Replicate weights
- Rotated test design
  - Plausible values
- Handling complex designs

Validity and testing

- The item pool should include a large number of items for domain validity (e.g., mathematical literacy)
- At the same time:
  - Fatigue biases results of long tests
  - Schools refuse to participate in lengthy studies
- Rotated test forms
  - Students are assigned a subset of item pool
  - Minimize testing time

Rotated booklet design

- Items are assigned into non-overlapping clusters
- Clusters of items are assigned into booklets
- Linking items (clusters) allow us to report on a single scale
  - Common items assigned to different students
- For example, in PISA 2003
  - 13 clusters
  - 4 clusters per booklet
  - 13 booklets

Rotated test forms

Table 5.9

PISA 2003 test design

	Block 1	Block 2	Block 3	Block 4
Booklet 1	C1	C2	C4	C10
Booklet 2	C2	C3	C5	C11
Booklet 3	C3	C4	C6	C12
Booklet 4	C4	C5	C7	C13
Booklet 5	C5	C6	C8	C1
Booklet 6	C6	C7	C9	C2
Booklet 7	C7	C8	C10	C3
Booklet 8	C8	C9	C11	C4
Booklet 9	C9	C10	C12	C5
Booklet 10	C10	C11	C13	C6
Booklet 11	C11	C12	C1	C7
Booklet 12	C12	C13	C2	C8
Booklet 13	C13	C1	C3	C9

Source: OECD (2009). *PISA Data Analysis Manual: SPSS (2nd Edition. Paris)*: OECD Publishing.

Plausible values

- Rotated booklets introduce challenges for estimating academic achievement
  - Students miss data on a number of items
- Plausible values methods are employed to obtain population estimates with rotated booklet designs
- Students do not answer all items but plausible scores are produced as if they had responded to all items based on
  - Responses to test items
  - Background characteristics

Plausible values

- Plausible values are random draws from the distribution of a student's ability
  - Instead of obtaining a point estimate, a range of values are estimated for each student
- A single score cannot be calculated because data is missing for a number of items
- Plausible values should not be used for individual performance
  - Plausible values should never be averaged at the student level

Imputation error

- Plausible values account for imputation error
  - Making inference on ability from small number of items
- Estimation should be conducted separately for each plausible value
  - Typically five plausible values are considered
  - The variability between estimates reflects the imputation error
- For example, for the correlation between SES and reading performance

Estimation

- Point estimate
  - Calculate the correlation separately for each of the 5 plausible values
  - Average the correlation
- Imputation error
  - Variability between estimates

Outline

- International assessment studies
- Two-stage sampling
  - Replicate weights
- Rotated test design
  - Plausible values
- Handling complex designs

Two sources of error

- Sampling error
  - Replicate weights
- Imputation error
  - Plausible values

Replicate weights and plausible values

- The calculation of unbiased estimates (e.g., mean, regression coefficient) and associated standard errors involves repeated computations of these estimates
- For example, 405 computations are needed in PISA
  - Sampling error:
    - 80 estimates (replicate weights) for each of the 5 plausible values (400 estimates)
  - Imputation error:
    - Each plausible value weighted by the total weight (5 estimates)

Replicate weights and plausible values

Table 8.1

The 405 mean estimates

Weight	PV1	PV2	PV3	PV4	PV5
Final	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
Replicate 1	$\hat{\mu}_{1\_1}$	$\hat{\mu}_{2\_1}$	$\hat{\mu}_{3\_1}$	$\hat{\mu}_{4\_1}$	$\hat{\mu}_{5\_1}$
Replicate 2	$\hat{\mu}_{1\_2}$	$\hat{\mu}_{2\_2}$	$\hat{\mu}_{3\_2}$	$\hat{\mu}_{4\_2}$	$\hat{\mu}_{5\_2}$
Replicate 3	$\hat{\mu}_{1\_3}$	$\hat{\mu}_{2\_3}$	$\hat{\mu}_{3\_3}$	$\hat{\mu}_{4\_3}$	$\hat{\mu}_{5\_3}$
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
Replicate 80	$\hat{\mu}_{1\_80}$	$\hat{\mu}_{2\_80}$	$\hat{\mu}_{3\_80}$	$\hat{\mu}_{4\_80}$	$\hat{\mu}_{5\_80}$
Sampling variance	$\sigma^2_{(\hat{\mu}_1)}$	$\sigma^2_{(\hat{\mu}_2)}$	$\sigma^2_{(\hat{\mu}_3)}$	$\sigma^2_{(\hat{\mu}_4)}$	$\sigma^2_{(\hat{\mu}_5)}$

Source: OECD (2009). *PISA Data Analysis Manual: SPSS (2nd Edition. Paris)*: OECD Publishing.

- Final mean estimate = average of final estimates
- Sampling variance = average of the variances
- Imputation variance = variance of final estimates
- Standard error = combined sampling and imputation variance

Problem

- Standard software cannot handle replicate weights and plausible values
- It is assumed that data were collected on a simple random sample
  - Standard errors are underestimated
  - Non-significant results tend to be significant
- Ignoring the complex design leads to wrong conclusions
  - Gender differences
  - Differences between countries

Available software

- IDB Analyzer (SPSS)
- NAEP Data Explorer (web tool)
- PISA SPSS macros
- R package 'intsvy'

R package 'intsvy'

- Free
  - Does not rely on commercial software like SPSS or SAS
- Open source
  - Can be extended to perform other analysis

Thank you!

daniel.caro@education.ox.ac.uk

C.Bokhove@soton.ac.uk