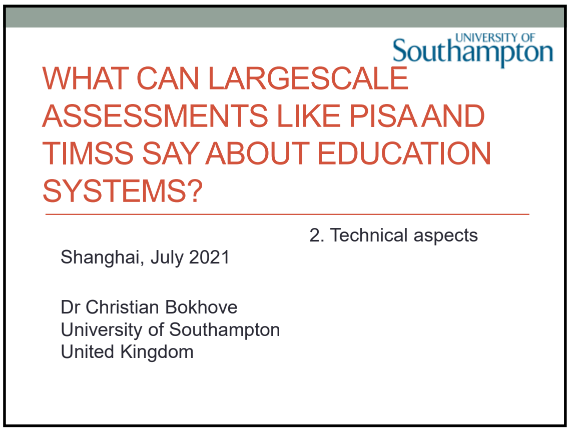


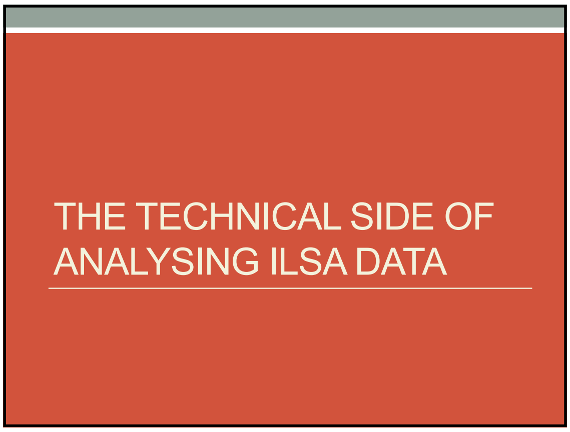
# WHAT CAN LARGESCALE ASSESSMENTS LIKE PISA AND TIMSS SAY ABOUT EDUCATION SYSTEMS?

2. Technical aspects

Shanghai, July 2021

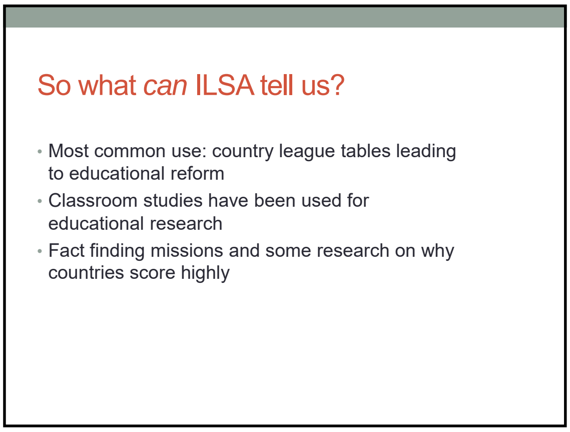
Dr Christian Bokhove
University of Southampton
United Kingdom

1



# THE TECHNICAL SIDE OF ANALYSING ILSA DATA

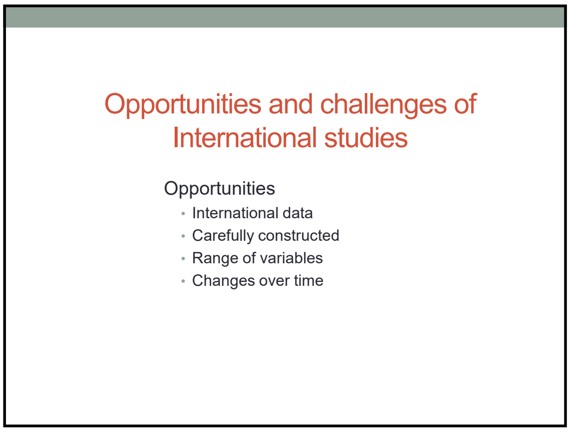
2



## So what *can* ILSA tell us?

- Most common use: country league tables leading to educational reform
- Classroom studies have been used for educational research
- Fact finding missions and some research on why countries score highly

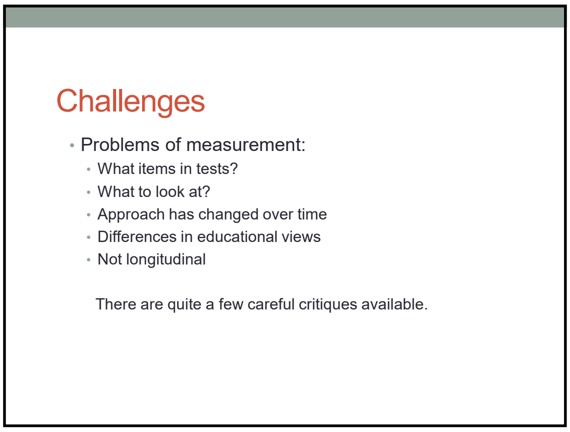
3



## Opportunities and challenges of International studies

Opportunities
- International data
- Carefully constructed
- Range of variables
- Changes over time

4

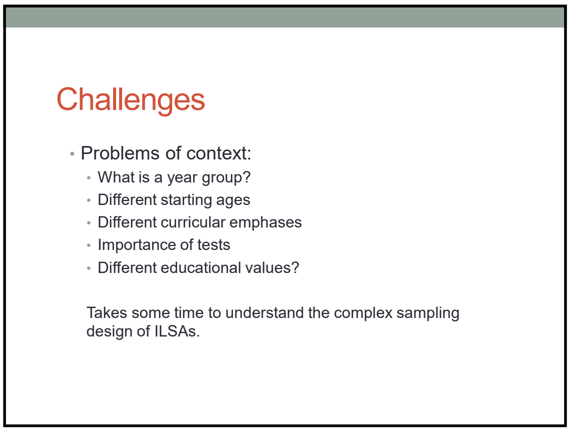


## Challenges

- Problems of measurement:
  - What items in tests?
  - What to look at?
  - Approach has changed over time
  - Differences in educational views
  - Not longitudinal

There are quite a few careful critiques available.

5



## Challenges

- Problems of context:
  - What is a year group?
  - Different starting ages
  - Different curricular emphases
  - Importance of tests
  - Different educational values?

Takes some time to understand the complex sampling design of ILSAs.

6

## Drawbacks of ILSAs

Chapter 6
**Extending Educational Effectiveness: A Critical Review of Research Approaches in International Effectiveness Research, and Proposals to Improve Them**

David Reynolds, Anthony Kelly, Alma Harris, Michelle Jones, Dennie Adams, Zhenzhen Miao, and Christian Bokhove

**6.1 Introduction: The Rise of International Effectiveness Research**

In the last 20 years considerable attention has been focused upon the variation between countries in educational achievement and the apparent relative effectiveness of their educational systems. Partly this is a reflection of a world that is, in many respects, becoming 'smaller' all the time. The spread of mass communications and new information technologies is affording all countries a more international 'reach' in their world views. The revolution afforded by the pervasive spread of information means that ideas now travel 'virally' around the world with great rapidity, making it increasingly possible for educational ideas and processes to move freely too. In education, the process of 'internationalisation' has taken longer to embed than in

https://www.dropbox.com/s/u4xf4ulf2yl36gq/chapter_eer_ilsa.pdf?dl=0

7

## 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

8

## 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

9

## 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)

10

## Dependency of observations

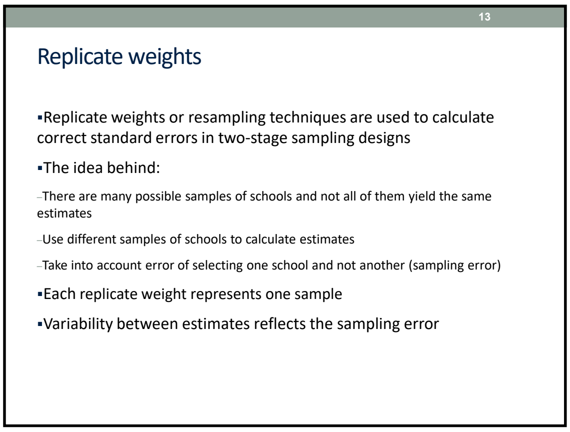
- Students within schools tend to be more similar than across schools
  - Family background
  - Instructional setting
- Observations within schools are not independent
  - Dependency of observations yields less information about the population
  - Uncertainty is greater in two-stage sampling
- Assume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

11

## 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

12

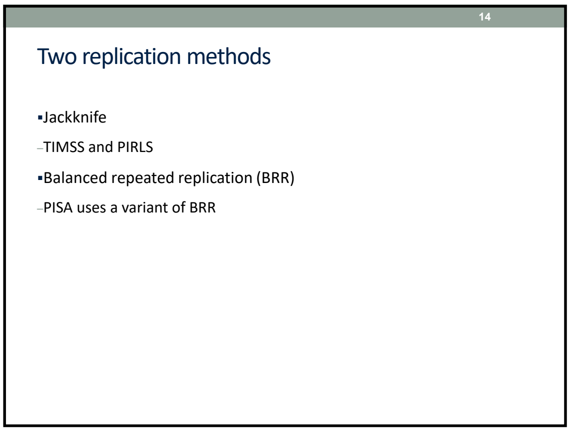


13

## 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

13

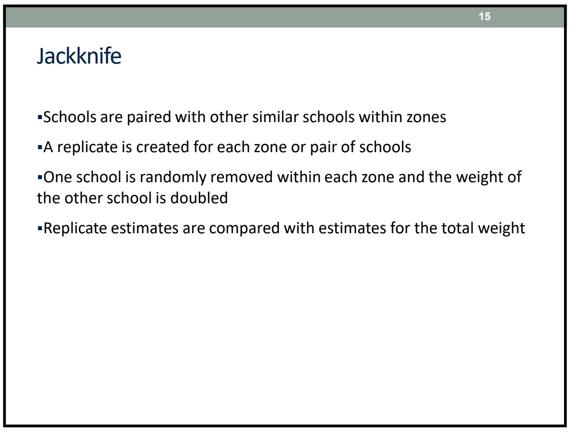


14

## Two replication methods

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

14

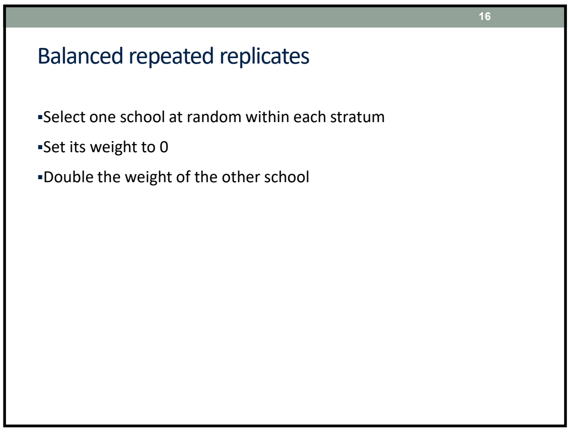


15

## 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

15

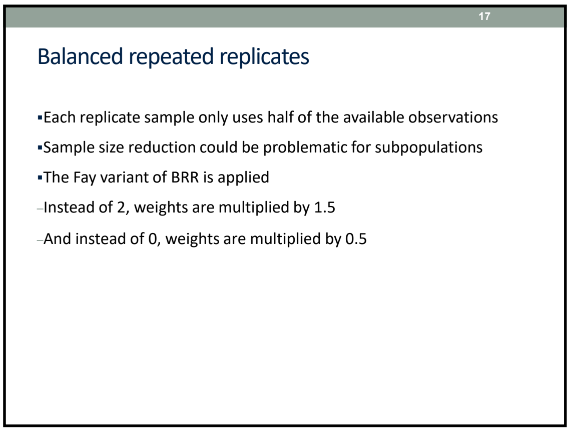


16

## Balanced repeated replicates

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

16

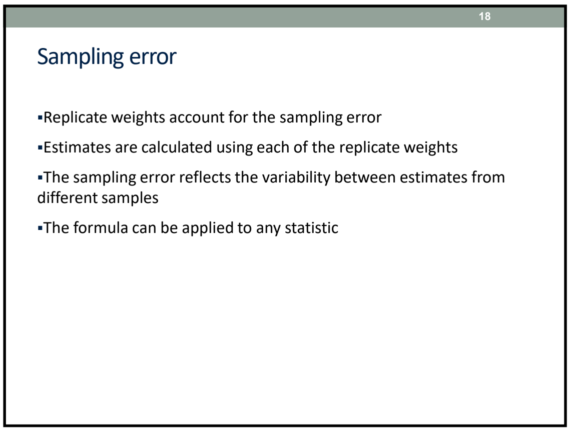


17

## 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

17



18

## 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

18

## 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

19

## 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

20

## Rotated test forms

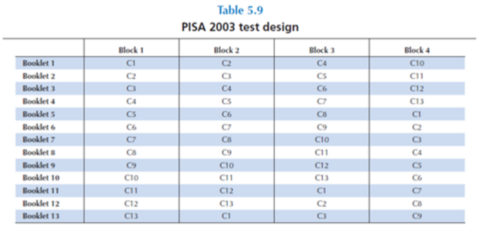


Table 5.9

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 (2$^{nd}$ Edition. Paris)*: OECD Publishing.

21

## 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

22

## 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

23

## 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, but for example PISA now uses ten
  - The variability between estimates reflects the imputation error
- For example, for the correlation between SES and reading performance

24

## Estimation

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

25

## Two sources of error

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

26

## 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)

27

## Replicate weights and plausible values

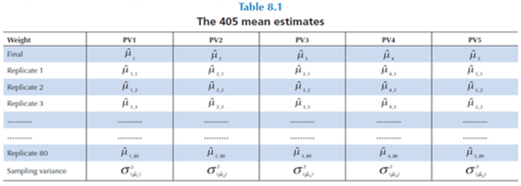


Table 8.1

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.

28

## Should not ignore complex sampling design

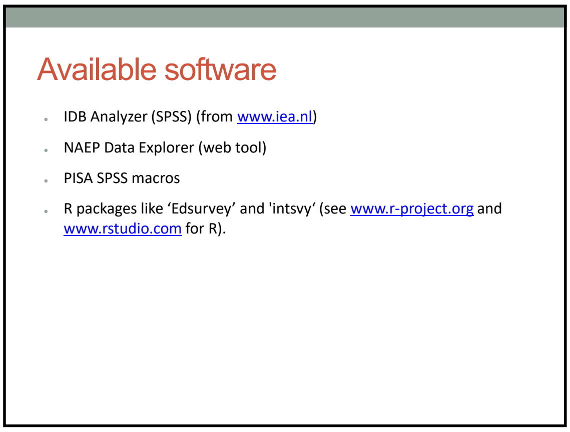
- Ignoring the complex design leads to wrong conclusions, like different point estimates and/or underestimated standard errors, see Rutkowski et al. (2010)
  - Variance estimation: jackknife, BRR
  - Not taking into account weights (e.g. Rutkowski et al (2010): Bulgarian TIMSS 2007, higher probability of selection to students from vocational and profiled schools). In a multilevel situation choosing wrong composite weights.
  - Treatment of plausible values: instead of Rubin's rules averaging (five) plausible values or choosing only one plausible value.
- Drent et al. (2013) formulated quality criteria (low, satisfactory, high)
- Standard software cannot handle replicate weights and plausible values

29

## 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

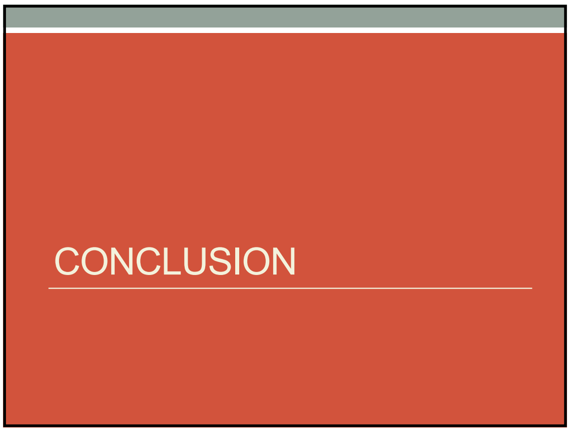
30



## Available software

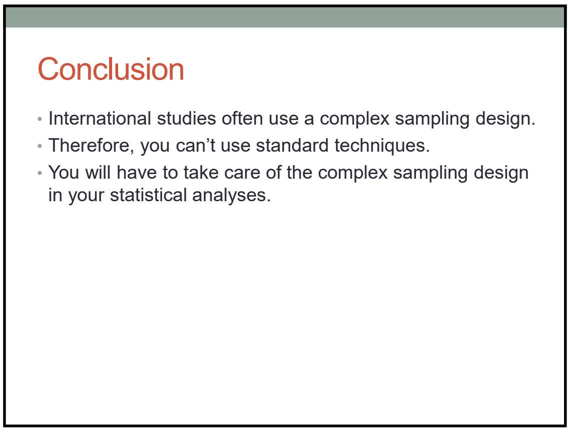
- IDB Analyzer (SPSS) (from www.iea.nl)
- NAEP Data Explorer (web tool)
- PISA SPSS macros
- R packages like ‘Edsurvey’ and 'intsvy‘ (see www.r-project.org and www.rstudio.com for R).

31



## CONCLUSION

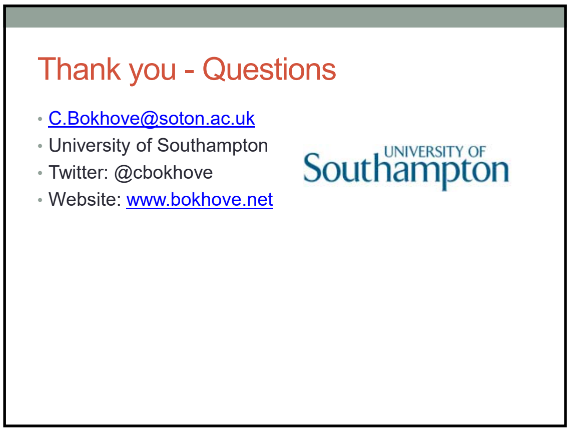
32



## Conclusion

- International studies often use a complex sampling design.
- Therefore, you can’t use standard techniques.
- You will have to take care of the complex sampling design in your statistical analyses.

33



## Thank you - Questions

- C.Bokhove@soton.ac.uk
- University of Southampton
- Twitter: @cbokhove
- Website: www.bokhove.net

34