Overview of the National Longitudinal Survey of Children and Youth
Methodologist

- Not subject-matter expert
- We do the sampling thing
- Determine the sample size
- How the sample is selected
- Analysis done by the methodologist
  - is to assess data quality
  - adjust estimation methods accordingly

Picking people for a sample, can be an expensive thing to do.

The job of a methodologist is to get the least expensive sample that will still produce the best estimates possible.

The methodologist helps determines the survey vehicle.
Goal of the Presentation

- The survey vehicle has an impact on the analysis.
- Today we’re going to describe the vehicle in order to facilitate the analytical process.
The Analytic

Being able to...

...make inferences

- partial non-response
- inconsistencies
- Sampling
  - Complex design
  - Impact on
    - estimation
    - precision
    - analysis
- Type of analysis
  - longitudinal, cross-sectional
  - other and mixed
NLSCY - Overview

The survey is almost entirely funded by HRDC
NLSCY - Overview

- Complex data structure
  - the lives of children are complex
  - dual child/household structure
  - new content in each cycle
  - some changes in old content

- Other constraints
  - limit on quantity of information
  - limited resources
Analytical Framework

Context

Resources

Family    School    Community    Work    Public programs

Social

Transitions:
- Health/injuries
- Accidents
- Mortality in family
- Periods of poverty
- Starting school
- Adolescence
- Graduation
- First job
- Marriage
- First child

Temporal

Time in years

Outcomes

- Physical health
- Emotion
- Language/communication
- Cognitive/learning
- Social

- Life-long learning
- Effective worker
- Effective parent
- Involved citizen
Unit of Analysis

- The child

- Sources of information
  - Person most knowledgeable about the child (PMK)
  - Teacher
  - School principal
  - Child himself/herself
    - cognitive measures
    - self-administered
Unit of Analysis

Caution

Other types of Analysis
  - Weights are designed for the child
  - Concepts like family are characteristics of the child
    - Not a domain for estimation

Statements like . . .

The NLSCY estimates the number of children with the family characteristic . . .
Cycle 3 Content
Respondents
PMK - Person Most Knowledgeable about the child. (May change over cycles)
Child/youth - Sampled child and Unit of analysis. (Constant over cycles)
Teacher(s)/School principal - associated with child’s education during cycle collection. ( Likely to change over cycles)

Child section answered by the PMK
- Education
- Health
- Medical and biological information
- Mother’s work after the child’s birth
- Child’s development
- Temperament
- Literacy
- Communication
- Activities
- Behaviour
- Positive behaviour
- Sleep habits
- Motor and social development
- Relationships
- Parenting
- Custody
- Expectations (Aspirations)
- Socio-demographic characteristics

Parents - answered by the PMK
- Education
- Labour force
- Income
- Health
- Social support
- Socio-demographic characteristics

Family - answered by the PMK
- demographics of members
- relationships within family
- family functioning
- dwelling
- Neighborhood safety

Self-completed (10-15 year-olds) - answered directly by the selected child / youth
- Friends and family
- School
- About me
- Feelings and behaviour
- My parents
- Puberty
- Smoking, drinking and drugs
- Activities

Not asked of 10-11 year-olds
- Dating
- Health
- Work and money

Direct Assessments
4-5 year olds
- Peabody Picture Vocabulary Test Revised (PPVT-R)
- Who Am I?
- Number Knowledge

Grades 2-10 (up to age 15)
- Math tests
- Reading tests*
  *(In Cycles 2 & 3 only)

Teacher - answered by single or multiple teachers
- teaching practices (s)
- Class demographics (s)

School - answered by school principal
- student population (s)
- disciplinary problems (s)
- school climate (s)
Products

- Master files (limited to STC and RDC users)
- Shared files (HRDC) – being phased out
- Public Use Micro Data files – Cycles 1 to 3
- Synthetic Data Files
  - Remote Access
  - Training
  - Off-site RDC users
- Custom tabulations
Time for maintenance

All vehicles must be followed up with regular maintenance.

The NLSCY is no exception.
With the release of Cycle 4 data, some changes to the previous cycles were done:

- Design weights
- Standardisation of FIELDRUK / PERSRUK
- MSD scores
- Health Utility Index
- Cycle 3 math scores
- Out-of scope omitted by mistake in Cycle 3
- Scales
Design weights

- Demographic estimates revised, using the latest Census information
  - Demographic estimates are used in the final stage of weighting (post-stratification).
  - NLSCY design for cycles 1 to 3 were revised accordingly
Design weights (cont’d)

- Cycle weights are final once the Census data are finalized.

- You can expect more revisions of NLSCY data.
Changes in the weights, changes in the variables and analysis ... Oh my!

- Changes in some longitudinal variables
  - improving the concept
  - different respondent
    - means a conceptual difference
  - different information about respondents
    - different PMK
    - response error
- Unprocessed responses
  - verification and consistency
  - partial responses
FIELDRUUK & PERSRUUK

- The identifiers:
  - FIELDRUUK (household Id)
  - PERSRUUK (person Id)

- Have been standardised for all cycles
  - Merging data from different cycles will be easier.
MSD scores

- Errors were found in the calculation of the MSD scores in past cycles.
  - Errors have been corrected
  - 2 derived scores are now available for each cycle
Health Utility Index

- In cycles 1 and 2, version 2 of the Health Utility Index (HUI) was used.
- In cycle 3, both HUI version 2 and version 3 are available.
- In cycle 4, only HUI version 3 is available.
Revised Cycle 3 maths scores

- An IRT (Item Response Theory) based score was omitted from the initial release (now included on the file).
- Mislabeled raw score was corrected as well.
Out-of-scope units omitted by mistake in Cycle 3

Deceased children and children who have moved out of the country were removed by mistake from the Cycle 3 master file:

- they are now included in the Cycle 3 master file
- Longitudinal weights adjusted, they have a cross-sectional weight of 0.
Scales

The NLSCY questionnaire has changed over the cycles (new questions added, some questions dropped).

- Impact on some scales. All the details are in the Cycle 4 Users’ Guide.
Cycle 4 Cognitive Tests and the Education Component

- Capacity problems in the field with STC surveys
  - Math test moved to home administration
  - Reading test was dropped
- Impossible to collect, for all kids, the school information in time to be at the school before the end of the school year.
  - Education file to be released in December is limited to a sub-sample
- School components was suspended for all but kindergarten kids in Cycle 5
In the NLSCY, you will find data are less processed.
Investment for the NLSCY

- Focus on derived variables
  - scales, cognitive measures
  - transition measures
- Non-response adjustment
  - total non-response
- Processing of financial data
  - family income
  - personal income
- Dissemination within reasonable time
Partial Non-response

- Respondent units are those which answered the key questions.
  - Not necessarily all the questions.
  - Some variables will include non-response, identified by:
    - not stated
    - don’t know
    - refusal
  - Sometimes an entire block is missing.
What Should You Do About Partial Non-response?

- Assess the extent of the partial non-response.
- Determine the impact on your analysis
  - By examining the variables related to the variable of interest
  - See if the missing responses can form a separate category
- Decide to do non-response processing
  - reweight for each variable to take partial non-response into account. Can be very tedious.
  - Document your processing procedures
Understanding Early Years continues to look at children aged 0 to 5

The longitudinal component continues to follow the children until the age of 25
Cross-sectional Analysis has been an important focus of analysis

Cross-sectional analysis will be limited to the UEY cohorts starting in Cycle 5
One Survey but actually many datasets

Intended for cohort analysis of 2 ages,
e.g., 0-1, 2-3, 4-5, 6-7, 8-9, 10-11
Dissecting NLSCY Data

Cross-sectional Data
Repeated Surveys

The sample size is very different from one cycle to the next, from one cohort to the next.

3 data cycles for children aged 0 to 11
2 data cycles for children aged 12 and 13

Early Years

0 1 2 3 4 15
1998-99

1994-95 0 110 11 2 132 130 10 1
2 1996-97
Issues

LONGITUDINAL ANALYSIS

- Limitations due to sample erosion
  - sample shrinkage problems
  - representation (coverage) problems
  - Swiss cheese problems

- Conditioning bias

- Interpretation of results
  - impact on effectiveness of estimation
  - inferences
Issues

CROSS-SECTIONAL ANALYSIS

- Limitations due to the age of the sample
  - Part of the sample was not selected for cross-sectional estimates
  - Inherent complexity in the sample design to meet divergent needs
  - Coverage problems
    - No update of the sample to reflect changes in the population (e.g., immigration); only the sampling weights have been adjusted to reflect changes
    - The older the cohort gets, the more difficult it is to adjust the sampling weights properly
Issues

**CROSS-SECTIONAL ANALYSIS**

- Limitations due to the nature of the survey
  - Some concepts were defined for the purposes of longitudinal analysis
  - Problems with sample erosion
  - Conditioning bias

- Interpretation of the results
  - Impact on the effectiveness of estimation methods
  - Making inferences
  - Greater potential with the supplementary samples that have been added
Issues

- MIXED ANALYSIS (longitudinal and cross-sectional)
  - Pay attention to the differences in the population targeted by the two types of analysis
  - Sample sizes vary a lot for these two types of analysis
  - Pay attention to the conclusions drawn from these analyses
  - The problems mentioned earlier can take different forms depending on the type of analysis
Dissecting NLSCY Data

Cross-sectional Data
Repeated Surveys

NOTE: The sample units are not independent of one another.
Issues

▲ CROSS-SECTIONAL ANALYSIS (REPEATED)

▲ Same limitations as noted earlier
▲ The sample overlaps from one cycle to the next.
▲ Independence or interdependence of samples
  □ There is sample interdependence when the sample is made up of the same respondents
  □ Involves a covariance factor
  □ Sample independence is possible only for certain domains (e.g., children aged 0-1)
One Survey but actually many datasets

An early childhood file

Early Years available in June
Dissecting NLSCY Data

Early Years
Cross-sectional Data

The sample size is very different from one cycle to the next, from one cohort to the next.

12,333 10,465 20,210

12,333 10,465 20,210

12,333 10,465 20,210

12,333 10,465 20,210

12,333 10,465 20,210

The sample size is very different from one cycle to the next, from one cohort to the next.

12,333
10,465
20,210

1,793
1,453
5,420

0 1 2 3 4 5

0 1 2 3 4 5

0 1 2 3 4 5

0 1 2 3 4 5

0 1 2 3 4 5

0 1 2 3 4 5

Or, once again, repeated analysis of the three data cycles

Or, once again, repeated analysis of the three data cycles

Or, once again, repeated analysis of the three data cycles

Or, once again, repeated analysis of the three data cycles

Or, once again, repeated analysis of the three data cycles

Or, once again, repeated analysis of the three data cycles

Or, once again, repeated analysis of the three data cycles
Analysis of Overlapping Domains

For example, analysis of children at a particular age

- There are actually 8,666 five-year-olds in the 3 cycles.
- The reference period becomes an analytical variable.
- Inference for a prescribed population is in context.

| Born in 1989 | 5 | 1,793 |
| Born in 1991 | 5 | 1,453 |
| Born in 1993 | 5 | 5,420 |

8,666
Issues

OTHER – OVERLAPPING ANALYSIS

▲ Pay attention to the definition of the subject area being studied

■ If the same respondent unit appears in the analytical sample more than once
  • Are they independent?
  • Do they impair the validity of the conclusions?

■ Relative weight of each unit in the analysis

■ Has the unit of analysis changed?

▲ Pay attention to the conclusions drawn from the analyses
Where Does the NLSCY Sample Come From?

Why ask the question?

- Issues concerning analysis
  - impact on statistical and analytical software applications
  - impact on subject areas being studied

Software tools:
- SAS
- STATA
- SPSS
- SUDAAN
Where does the sample come from?

Most of the sample is taken from the Labour Force Survey out-rotating sample:
- Sometimes active rotations have to be used for 0 to 1 year-olds.
- 12 rotation groups can provide close to 2000 kids of a particular age (one calendar year of LFS samples).
- So the information is never more than a year old.

When very large samples are needed (more than 2000 for a particular age) then the Birth Registry is used:
- The information is at least 1 year old
- For certain ages (i.e., 5 year-olds) the information can be quite old
The NLSCY Sample

A large part of the sample comes from the LFS

- geographic stratification
- multi-stage, with the primary sampling unit being a geographic cluster
- with a systematic sample of households
- After the LFS interview, households are identified as containing or not containing the population of interest (children)
- The in-sample unit is the child (not the household)

- Constitutes the initial frame of children selected in 1994
- Main source of newborns and cross-sectional samples
Applications such as SAS and SPSS can use sampling weights for estimation purposes, but they do not use them correctly for variance estimation. Some applications can use sampling weights for estimation purposes and can estimate the variance correctly. But they cannot take into account the NLSCY’s special problems.
Where Does the NLSCY Sample Come From?

− Other reasons for asking the question
  − Issues concerning the type of analysis that can be done
    • cross-sectional, longitudinal, other
    • Stratum jumpers
  − Issues concerning interpretation of results
    • impact on the effectiveness of estimation methods
    • inferences
Implications for Analysis

- The structure favours analysis for geographic areas
- Loss of effectiveness for other subject areas
- The advantage is reflected in a gain in operational efficiency
- Can use a larger sample for the same cost
- Target the analysis to take advantage of this structure
- Some estimation methods can improve efficiency
The NLSCY Sample

- Sample size changes substantially from one initiative to another
  - between the Early Years (EY) cohorts and the longitudinal cohort
  - from one cycle to the next
  - because the sample size was increased with exclusively cross-sectional units.
The NLSCY Sample

- Renewal of the sample with EY cohorts is not consistent
  - Focus is on new age 0-1 cohorts
  - Specific additions such as EY (five-year-olds)
- This is a consequence of having analytical goals that vary from one initiative to the next and one cycle to the next
**The NLSCY Sample**

- For large supplementary samples
  - we used the Birth Registry
    - geographic and demographic stratification
    - simple random sample of geographic clusters
    - with a subsample of children in each cluster
  - it’s the option used for the samples of one-year-olds and five-year-olds in Cycles 3 and 4.
Link between analysis and the sample design

- Child’s Ability
- Intelligence
- Social environment
- Teachers
- School
- Materials
- Curriculum
- Grade level
- Subject
- Province

The proportion of kids in the sample being taught the PEI curriculum is much larger than what’s found in the population.

Province is a stratum
A Key Goal of Analysis...

- Be able to derive facts from the data
- A scientific sample design allows for:
  - estimates
  - inferences
  - a degree of certainty
- The analysis is judged on its ability to:
  - back up its statements
  - prove its validity
We have seen that the survey is loaded with information
We can lighten the load by targeting our analysis
The Survey Vehicle

We can see what’s possible and what’s not
The Survey Vehicle

We can greatly improve the survey’s effectiveness by taking advantage of the way it’s constructed.
When we adjust our analytical approach, the vehicle becomes lighter and more manoeuvrable.
The Survey Vehicle

And we know we’re not alone doing analysis
The Survey Vehicle

NLSC

Questions
How to find stuff about the NLSCY

www.statcan.ca/english/ssds/44500.htm
www.statcan.ca/english/ips/data/89f0078x1e1999003.htm

www.hrdc-drhc.ca/sp-ps/arb-dgra/nlscy-elnej/home.shtml
My coordinates

André Cyr
Room 2500, Main Building, Statistics Canada
Tunney`s Pasture, Ottawa, Ontario
K1Y 4P2

andre.cyr@statcan.ca
Tel: (613) 951-1452