Use of crowdsourcing at Statistics Canada during the COVID-19 pandemic

Presentation to the Centre interuniversitaire québécois de statistiques sociales

Martin Renaud
March 30, 2021

Delivering insight through data for a better Canada

Presentation plan

1. Background
2. Methodology's role
3. Methodological considerations
4. Research
5. Moving forward

1. Background

- Crowdsourcing: A non probabilistic method of collecting data by inviting all members of a target population to voluntarily participate in a data collection exercise on a topic of interest.

- Previous crowdsources in Statistics Canada
  - Building register (Ottawa-Gatineau area)
  - Cannabis
  - Market basket measure

- Document on methodological considerations on crowdsourcing

1. Background

- March 2020: COVID-19 pandemic
  - Data gap ➡ urgent need for timely data
  - Country wide shutdown
  - Citizens engagement
  - Favorable conditions to use crowdsource as a data collection tool

➡ **Go ahead** ➡ **Impact of COVID-19 on Canadians: Data Collection Series**

1. Background

9 completed crowdsourcing projects

Impacts of COVID-19 on Canadians, April 3 to April 24

Postsecondary students, April 19 to May 1

Your Mental Health, April 24 to May 11

Perceptions of Safety, May 12 to May 25

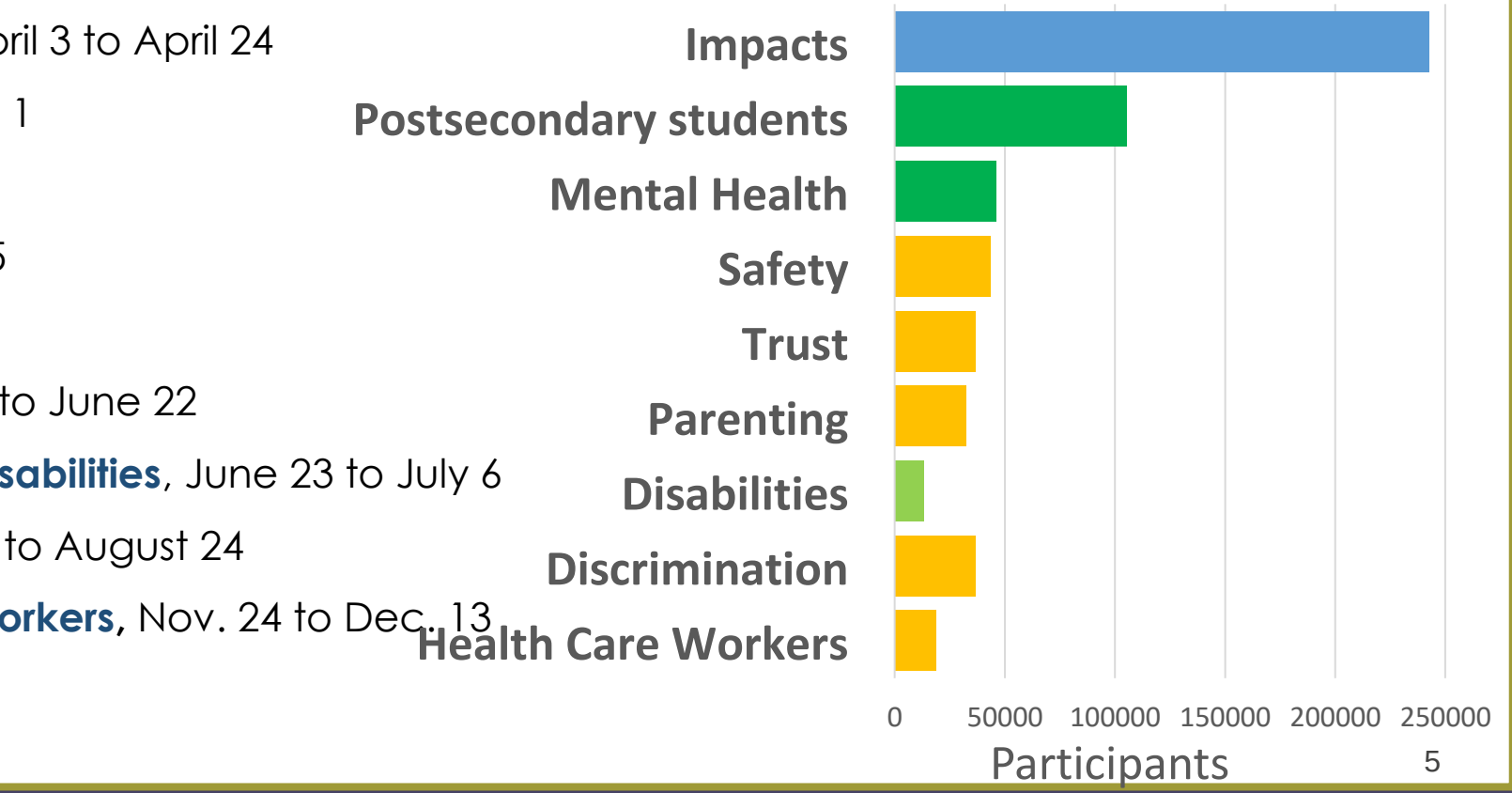
Trust in Others, May 26 to June 8

Parenting During the Pandemic, June 9 to June 22

Living with Long-term Conditions and Disabilities, June 23 to July 6

Experiences of Discrimination, August 4 to August 24

Impacts of COVID-19 on Health Care Workers, Nov. 24 to Dec. 13



1. Background

- Email address & phone number of participants asked at every crowdsource.
  - Used to contact participants in upcoming initiatives.
  - A little more than 50% provide an email address (almost 130,000 addresses after 1st crowdsource).

- Previously collected email addresses used to stimulate participation.
  - Started with Mental Health crowdsource.
  - Used only when applicable (ex: not used for parenting crowdsource)

- Maintenance and management of email database

2. Methodology's role

- Participation rate << 1% ➔ Bias

- Efforts geared towards
  - Mitigating bias
  - Ensuring rigor and data quality
  - Validating results
  - Guiding interpretation and analysis

2. Methodology's role

- Mitigating bias
  - 3 of 4 participants are women
  - Teens and seniors, low participation
  - Ontario and Nova Scotia participate well, Quebec not so much
  - Daily collection monitoring reports
    - Targeted follow-up

Gender	Age Group	Province						Total (missings incl.)
		Atlantic	Quebec	Ontario	Prairies	BC	Terr.	
Male	1	18	87	124	38	31	0	299
	2	157	480	860	281	224	14	2,018
	3	232	476	1,204	394	315	13	2,639
	4	189	410	977	305	249	8	2,143
	5	153	318	870	253	233	10	1,846
	6	90	276	647	175	225	8	1,425
	Total (missings incl.)	840	2,048	4,690	1,448	1,281	53	10,392
Female	1	59	75	229	87	55	1	510
	2	414	716	2,221	716	564	31	4,671
	3	728	1,064	3,489	1,194	1,089	42	7,616
	4	670	891	2,636	768	847	39	5,865
	5	478	661	2,006	670	709	36	4,565
	6	259	312	1,235	364	525	19	2,725
	Total (missings incl.)	2,613	3,723	11,837	3,804	3,793	168	25,998
Total (missings and others included)	1	85	165	370	138	90	1	855
	2	594	1,217	3,148	1,031	815	45	6,863
	3	967	1,548	4,738	1,615	1,424	55	10,366
	4	867	1,306	3,636	1,081	1,110	47	8,067
	5	635	982	2,890	929	945	46	6,442
	6	350	591	1,886	542	758	27	4,170
	Total (missings incl.)	3,504	5,815	16,700	5,346	5,151	222	36,851

2. Methodology's role

- Mitigating bias
  - Benchmarking strategy (benchmarking factors)
  - Basic adjustments to known control totals (ex: sex/age groups/geography)

$$BF_{ijk} = \frac{N_{ijk}}{n_{ijk}}$$

i: sex *j: age group* *k: province/territory*

$$\sum_{i,j,k} N_{ijk} = N$$

$$BF^{*}_{ijk} = \frac{n}{N} BF_{ijk}$$ → To be used on the Public Use Microdata File

  - Collapsing as necessary

2. Methodology's role

- Mitigating bias: control totals used in benchmarking
  - General population, Mental health, Perceptions of safety, Trust in others
    - February 2020 demographic projections, number of people by province, sex, age groups
  - Postsecondary students
    - Projected number of students enrolled in a postsecondary program on March 1, 2020 by sex, province of study, level of study

2. Methodology's role

- Mitigating bias: control totals used
  - Parenting during a pandemic
    - 2020 projected number of families by province with children
      - 0-5 years old / 6-14 years old / 0-14 years old
  - Living with long-term conditions and disabilities
    - 2016 Census counts of people living with a long-term condition or a disability by province, sex, age group

2. Methodology's role

- Mitigating bias: control totals used
  - Experiences with discrimination
    - June 2020 demographic projections, number of people by province, sex, age group, and visible minority status
  - Health care workers
    - Numbers of health care workers by province and type of job

2. Methodology's role

- Ensuring rigor and data quality
  - Basic data verifications
    - Ex: Valid postal code
  - Outlier detection
    - Ex: Unusually large number of children in household
  - Out of scope records
    - Ex: Individuals less than 15 years old
  - Collection issues
    - Ex: Multiple answers from one participant

2. Methodology's role

- Validating results
  - Comparisons to other reliable sources when possible
    - Ex: Crowdsource #1 vs Canadian Perspective Survey Series #1
    - Ex: Gender diverse vs Census Test results
  - Raw vs benchmarked results

2. Methodology's role

- Guiding interpretation and analysis
  - Adapted terminology
    - Respondents ➔ Participants
    - Survey ➔ Data collection initiative
    - Estimates ➔ Results
    - Weighting ➔ Benchmarking
  - No data quality measures
    - CV's, confidence intervals, margin of error

2. Methodology's role

- Guiding interpretation and analysis
  - Warning about crowdsourcing

> *Unlike other surveys conducted by Statistics Canada, crowdsourcing data are not collected under a probability sampling design. As a result, the findings reflect only the responses of those who completed a questionnaire, and thus cannot be generalized to the entire Canadian population.*

  - Proportions only, no totals

3. Methodological considerations

- Document on methodological considerations about data collection using crowdsourcing in Statistics Canada (SSMD, April 2020)
  - Context
    - What is crowdsourcing?
    - Purpose of crowdsourcing
  - Considerations
    - Inferential limitations
    - *Preparing for crowdsourcing data collection*
    - *Releasing results from crowdsourcing*

3. Methodological considerations

- Preparing for crowdsourcing data collection
  - Demonstration of the necessity of collecting such data (principles of necessity and proportionality)
  - Management approval
  - Questionnaire approval
  - Consultation with:
    - Subject matter
    - Methodology
    - Data collection experts
    - Communications experts

3. Methodological considerations

- Preparing for crowdsourcing data collection
  - Development of the crowdsourcing tool
    - Data verification rules (outlier prevention)
    - Quality indicators summarizing verifications and assessing comparability
    - Safety measures:
      - In-scope IP addresses
      - High number of responses from one IP address
      - Prevention against bot answers
  - Outlier detection
  - Confidentiality measures

Identified as a priority by the Advisory Committee on Statistical Methods

3. Methodological considerations

- Releasing results from crowdsourcing
  - Results should be accompanied by the relevant following statements to inform users of the data limitations
    - Who was invited to participate on a voluntary basis, and how
    - Number of valid questionnaires
    - Start and end of data collection
    - Information about outlier detection, data correction, data exclusion
    - Disclaimer about the absence of a sampling design → No measures of precision
    - Information about benchmarking process
    - Warning about bias and interpretability
    - Quality indicators produced

3. Methodological considerations

- In the process of developing a directive on crowdsourcing at Statistics Canada
  - Legal context
  - Definitions
  - Objective and expected results
  - Responsibilities of every party involved

4. Research

Why was research needed?

- First crowdsourcing cycles:
  - No time to develop complex methods for reducing bias
  - Implemented post-stratification by age, sex and province:

    post-stratified estimates close to unweighted estimates

- Need to use other auxiliary variables and more sophisticated weighting methods for further bias reductions
- Four research projects on data integration methods: combining crowdsourcing data with data from other probability surveys

4. Research

Objectives

1. Reduce the bias of the crowdsourcing estimates
   - Propensity score weighting
     - Some auxiliary variables (ex: education) are effective at reducing bias
     - Significant amount of bias still remains
   - Sample matching
     - Not as effective as propensity score weighting to reduce bias

4. Research

Objectives

2. Reduce the variance of probability survey estimates

- Small area estimation
  - Substantial precision gains are observed
  - Not for all characteristics of interest (smaller for proportions near 0 or 1)
- Dual frame weighting
  - Variance reduction is quite small in general
  - Larger variance reduction expected with a moderate to large participation rate

5. Moving forward

- Careful analysis of the needs before using crowdsourcing
- What is important? Time, cost, accuracy, …?
- Other options
  - Alternative sources of data (if available)
  - Probability surveys
    - Better control over the potential bias even when the sample size is small
    - Similar time frame in some cases
    - Ex.: Canadian Perspective Survey Series

5. Moving forward

- Continue research projects
- Develop quality indicators
- Protect against vulnerabilities
- Finalize directive on crowdsourcing

QUESTIONS?

Thank you!

Martin Renaud

martin.renaud@canada.ca