User Guide for the Crowdsourcing: Impacts of the COVID-19 on Canadians – Experiences of discrimination

Public Use Microdata File (PUMF)

October 2020

1. Introduction

The data collection series Crowdsourcing: Impacts of COVID-19 on Canadians is designed to assess the quality and viability of a more timely collection model using willing participants and web-only collection. The Crowdsourcing: Impacts of COVID-19 on Canadians – Experiences with discrimination is the seventh iteration in the continuing series of crowdsourcing cycles. The overall goal of the crowdsourcing initiative is to invite all members of the Canadian population to participate in a data collection exercise on a voluntary basis. The main topic of this seventh crowdsourcing was to evaluate how the COVID-19 pandemic impacted confidence and trust in various institutions, general public, and neighbours, and to determine if experiences with discrimination before and during the pandemic has disproportionally impacted certain groups more than others.

In the context of this product, the term *crowdsourcing* refers to the process of collecting information via an online questionnaire. Open advertising was used to obtain participants who chose to self-select by completing the questionnaire. As such, the crowdsourcing data was collected through a completely non-probabilistic approach which does not involve a random selection of respondents like other traditional Statistics Canada surveys. Therefore, results pertain only to the participants and cannot be used to draw conclusions about the larger population of individuals in Canada who live with a long-term condition or disability.

The following sections of the document provide a summary of different methodological considerations relevant to this crowdsourcing as well as information on how the Public-Use Microdata File (PUMF) was created, how it should be used, and what its limitations are.

2. Confidentiality

Statistics Canada is prohibited by law from releasing any information it collects that could identify any person, business, or organization, unless consent has been given by the respondent or as permitted by the Statistics Act. Various confidentiality rules are applied to all data that are released or published to prevent the publication or disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data.

The approach for creating this PUMF is to balance the requirements for maintaining the confidentiality of participants by minimizing disclosure risks, while providing the most useful data to users. The production of a PUMF includes many safeguards to prevent the identification of any one person. Confidentiality of the participants to the crowdsource is ensured mainly by the reduction of information. Some variables that were collected (ex: postal code) do not appear on the PUMF due to suppressions (see Appendix A) while response categories have been restricted or collapsed for other variables in order to reduce disclosure risk (see Appendix B).

3. Methodology

3.1 Collection period

Collection for this crowdsourcing started on August 4, 2020 and closed on August 24.

3.2 Target population

The target population for this crowdsourcing initiative was all Canadians aged 15 and up, living in one of the ten provinces or three territories during the collection period.

3.3 Data sources and collection tool

Participation in this crowdsourcing initiative was voluntary. Prompts to participate were done through social media as well as a variety of outside partners like other government agencies, private and public organizations, associations, and news channels. Data were collected directly from participants via a self-administered online questionnaire found on an anonymous portal on Statistics Canada's website (i.e. the crowdsourcing application). Data collection was available in English and in French and the questionnaire took approximately ten minutes to complete. The questionnaire followed standard practices and wording used in a computer-assisted interviewing environment, such as the automatic control of flows that depend upon answers to earlier questions and the use of edits to check for logical inconsistencies and capture errors. The computer application for data collection was tested extensively.

3.4 Verifications

The following validation rules were implemented during processing of the data to comply to the target population of the crowdsourcing.

- If the age of the participant was missing or less than 15 years old, the record was set as out of scope.
- If sex at birth was missing, the record was set as out of scope.
- Participants who reported a gender of "Other" were coded where possible using their write-in response to "Male", "Female", "Other", or "Invalid".
- Participants who reported a visible minority status of "Other" were coded where possible using their write-in response to one of the eleven valid responses, "Not Included Elsewhere" or "Invalid".
- If the postal code was missing, started with an invalid letter (not assigned to a Canadian province or territory), did not have a number in the second character, had the following first three digits (A9A, H0H, N1N, X0X, X1X), or was A0A 0A0 or N0N 0N0, the record was set to out of scope.

No imputation was performed and questions that were left unanswered by participants should be excluded from analysis. For each question, only a very small percentage of participants (less than 2%) did not provide an answer.

3.5 Sample design

Crowdsourcing is a non-probabilistic approach to collecting data which does not use a sample design. Unlike probability-based surveys which select a sample of units using a controlled random mechanism, crowdsourcing participants provide their information on a voluntary basis. They are not sampled with a known probability of selection and therefore, a survey weight cannot be calculated.

3.6 Coverage

In the absence of a sample design, the coverage of the population cannot be measured. It is possible that some participants who provided their data were not actually part of the target population. It is also possible that some participants completed more than one questionnaire. In the crowdsourcing context, it is acceptable to have individuals participate more than once during the collection period. This could be the case if their opinion has changed for example. Verifications were performed to detect an abnormally large number of responses from one person and none were found.

Canadians not in tune with the various channels used to advertise the crowdsourcing as well as individuals with lower propensity to participate in surveys and data collection exercises are not well represented by the collected data. This translates into over/underrepresentation for some groups of the population. For example, males, youth (15 to 24 years old), seniors (65 years and older), and people from Quebec, Saskatchewan, and Alberta were underrepresented. People who reported a visible minority were also underrepresented. On the other hand, women, adults 35 to 54 years old, as well as people from Prince Edward Island, Nova Scotia, New Brunswick, Ontario, and Yukon were overrepresented. People who did not identify as a visible minority were also overrepresented. To generalize results observed from the participants to the population, one would have to make the assumption that participants to the crowdsourcing are representative of those not participating, which is an assumption that cannot be validated in practice.

3.7 Benchmarking

Benchmarking the crowdsourced data to known population totals can, under certain circumstances, partly reduce some of the self-selection and coverage bias but cannot adequately correct for it completely. To compensate for the over/underrepresentation of the participants, benchmarking was done in a similar way as calibration or post-stratification is done in a probability-based survey. In this case, benchmarking helped to correct for differing participation rates across province/territory, sex, age group, and categories of visible minority.

Demographic projections of the number of people by province, sex, age group and visible minority status as of June 2020 were used as control totals to calculate a benchmarking factor for every participant. The initial goal was to benchmark exactly by province and sex for the following age groups: 15-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+ with an additional calibration adjustment to match visible minority population projections at the national level within a certain tolerance. However, some collapsing of those age groups had to be done because of the small number of participants in some of them. The final collapsing strategy used to produce the benchmarking factors was as follows:

- No collapsing needed for either sex in New Brunswick, Quebec, Ontario and Manitoba, or for women in Newfoundland and Labrador, Nova Scotia, Saskatchewan, Alberta, and British Columbia.
- Collapsing of men in Newfoundland and Labrador, Nova Scotia, Saskatchewan, Alberta, and British Columbia, as well as women in Prince Edward Island to the following three age groups: 15-34, 35-49, 50+.
- Collapsing men of Prince Edward Island to the following two age groups: 15-44, 45+.
- Yukon, the Northwest Territories and Nunavut were grouped into a single region. Each sex was collapsed to the following three age groups: 15-34, 35-49, 50+.

The benchmarking factors were standardized in a way that their sum totals up to the number of participants instead of the population size in each benchmarking group.

4. Guidelines to analysis

4.1 Benchmarking factors

The microdata on the PUMF are unweighted. It is the responsibility of data users to apply the standardized benchmarking factors in any results they wish to produce. Benchmarking factors for this crowdsourced data were calculated to correct for differing participation rates across province/territory, sex, age group, and visible minority category. If these factors are not used, the results derived from the microdata will not correspond to those that would be produced by Statistics Canada.

Standardized benchmarking factors should be used to produce results in the same way weights are used to produce estimates from a probabilistic survey. **However, because of the non-probabilistic nature of crowdsourcing and the calculation of a standardized benchmarking factor, results should be limited to proportions only and data from the crowdsourcing should not be used to calculate totals**. Furthermore, results cannot be used to draw conclusions about the Canadian population.

4.2 Data quality indicators

Given the non-probabilistic nature of the crowdsourcing data collection and the absence of a sample design, a probability of selection and a survey weight are not available for this product. Even though a standardized benchmarking factor has been calculated, it should not be used to calculate measures of precision commonly associated with probabilistic surveys (e.g. coefficients of variation, margins of error, confidence intervals) as the results would not be valid.

4.3 Rounding guidelines

It is strongly recommended that users adhere to the following guidelines regarding the rounding of results produced from crowdsourced data.

a) Proportions and ratios are to be computed from unrounded components (i.e. numerators and/or denominators) and then are to be rounded themselves to three decimals using normal rounding. In normal rounding to a single digit, if the final or only digit to be dropped is 0 to 4 the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is increased by 1.
b) Rounding of totals and differences in aggregates does not apply as crowdsource data should only be used to calculate proportions.
c) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used, with the result that results to be published or otherwise released differ from corresponding results published by Statistics Canada, users are strongly advised to indicate the reasons for the differences in the documents to be published or released.
d) Under no circumstances are unrounded results to be published or otherwise released by users. Unrounded results imply greater precision than actually exists.

4.4 Minimum sizes

For reliability purposes in the calculation of proportions, it is recommended that the numerator should have at least 10 participants displaying the characteristic of interest while the denominator should have at least 30 participants in the domain of interest. Given the non-probabilistic nature of crowdsourcing, the user should not publish proportions of 0% or 100%.

4.5 Other considerations

a) Given the non-probabilistic nature of crowdsourcing, the user is reminded that the data cannot be used to draw conclusions for the Canadian population. The smaller the domain of interest is, the more likely the results are to be biased and not representative. The user is advised to proceed with extreme caution.
b) Given the rapidly evolving situation with regards to the COVID-19 pandemic, appropriate context should be provided about the data collection period as well as the nature and extent of restrictions in place at that time when reporting results.
c) The content of the questionnaire asked participants to reflect how they felt at the time of collection. This should be made clear when reporting results.

Appendix A

The following is the list of variables removed from the file in the creation of the PUMF.

Variables removed

Variable name	Description
DEM1_30A	Born in Canada
DEM1_30B	Canadian citizen
DEM1_30C	Landed immigrant or permanent resident
DEM1_35	Year where the participant became a landed immigrant/permanent resident
GENDER	Gender of participant (3 category)
DEM_05	Age of participant
AGEGR_5	Age group in increments of 5
AGEGR_10	Age group in increments of 10
DEM_13	Household size
DEM_14	Number of children (≤ 18 years old) in the household
DEM_15	Postal code
CMA	Census metropolitan area
CT	Census tract
CSDUID	Census sub-division
UCE_01	Sign-up for future survey
UCE_05C	Phone number
IIDENT	Indigenous identity group
IMMYR	Number of years in Canada since immigration
VISMIN	Visible minority group
RESPLANG	Language of submission

Appendix B

Recoded variables

<table>
<tr><th>Variable Description</th><th>On Master File</th><th>On PUMF</th></tr>
<tr><td>Age group</td><td>Variable name: AGEGROUP
1 15 to 24 years old
2 25 to 29 years old
3 30 to 34 years old
4 35 to 39 years old
5 40 to 44 years old
6 45 to 49 years old
7 50 to 54 years old
8 55 to 59 years old
9 60 to 64 years old
10 65 years old and over</td><td>Variable name: PAGEGR
Combined age groups based on the ones used for benchmarking. Also based on AGEGROUP, classes are combined according to province and sex.
1 15 to 34 years old
2 35 to 44 years old
3 45 to 54 years old
4 55 years old and over
5 15 to 34 years old
6 35 to 49 years old
7 50 years old and over
8 15 to 44 years old
9 45 years old and over
10 15 years old and over

Specific code set by province and sex are used:
<table>
<tr><th rowspan="2">Province</th><th colspan="2">PAGEGR available</th></tr>
<tr><th>Sex = 1</th><th>Sex = 2</th></tr>
<tr><td>10</td><td>8,9</td><td>1,2,3,4</td></tr>
<tr><td>11</td><td>10</td><td>8,9</td></tr>
<tr><td>12</td><td>8,9</td><td>1,2,3,4</td></tr>
<tr><td>13</td><td>1,2,3,4</td><td>1,2,3,4</td></tr>
<tr><td>24</td><td>1,2,3,4</td><td>1,2,3,4</td></tr>
<tr><td>35</td><td>1,2,3,4</td><td>1,2,3,4</td></tr>
<tr><td>46</td><td>1,2,3,4</td><td>1,2,3,4</td></tr>
<tr><td>47</td><td>8,9</td><td>1,2,3,4</td></tr>
<tr><td>48</td><td>8,9</td><td>1,2,3,4</td></tr>
<tr><td>59</td><td>8,9</td><td>1,2,3,4</td></tr>
<tr><td>60, 61, 62</td><td>5,6,7</td><td>5,6,7</td></tr>
</table></td></tr>
<tr><td>Province</td><td>Variable name: PROV
10 Newfoundland and Labrador (NL)
11 Prince Edward Island (PE)
12 Nova Scotia (NS)
13 New Brunswick (NB)
24 Quebec (QC)
35 Ontario (ON)</td><td>Variable name: PPROV
10 Newfoundland and Labrador (NL)
11 Prince Edward Island (PE)
12 Nova Scotia (NS)
13 New Brunswick (NB)
24 Quebec (QC)
35 Ontario (ON)</td></tr>
</table>

<table>
<tr><td></td><td>46 Manitoba (MB)
47 Saskatchewan (SK)
48 Alberta (AB)
59 British Columbia (BC)
60 Yukon (YT)
61 Northwest Territories (NT)
62 Nunavut (NU)</td><td>46 Manitoba (MB)
47 Saskatchewan (SK)
48 Alberta (AB)
59 British Columbia (BC)
63 Territories (YT, NT and NU)</td></tr>
<tr><td>CMA/CA Size</td><td>Variable name: CSizeMIZ
1 1,500,000 +
2 500,000 - 1,499,999
3 100,000 - 499,999
4 10,000 - 99,999 (any CMACA < 100,000)
5 Non-CMA/CA; Strong MIZ
6 Non-CMA/CA; Moderate MIZ
7 Non-CMA/CA; Weak/No MIZ, Territories outside of any CA
9 Missing</td><td>Variable name: PCSizMIZ
1 1,500,000 +
2 500,000 - 1,499,999
3 100,000 - 499,999
4 10,000 - 99,999 (any CMACA < 100,000)
5 Non-CMA/CA
6 10,000 - 499,999
7 10,000 - 99,999 (any CMA/CA < 100,000) or Non-CMA/CA
9 Missing

The available code set depends on the province:<table><tr><td>Provinces</td><td>Codes available</td></tr><tr><td>11, 12, 13, 24, 35, 47, 48, 59</td><td>1, 2, 3, 4, 5</td></tr><tr><td>10, 46</td><td>1, 2, 3, 7</td></tr><tr><td></td><td>1, 2, 6, 5</td></tr><tr><td>63</td><td>9</td></tr></table></td></tr>
<tr><td>Education Level</td><td>Variable name: ED_05
1 Less than high school diploma or its equivalent
2 High school diploma or a high school equivalency certificate
3 Trade certificate or diploma
4 College/CEGEP/other non-university certificate or diploma
5 University certificate or diploma below the bachelor’s level
6 Bachelor’s degree
7 University certificate, diploma, degree above the BA level</td><td>Variable name: PED_05
1 Did not attend university
2 Attended university</td></tr>
<tr><td>Language(s) used at home</td><td>Variable name: LAN_02A
1 English</td><td>Variable name: PLAN_02</td></tr>
</table>

	Variable name: **LAN_02B** 2 French Variable name: **LAN_02C** 3 Other language	1 Both English and French are mostly used at home 2 English is mostly used at home 3 French is mostly used at home 4 Other language(s) is (are) mostly used at home
Marital Status	Variable name: **DEM_11** 1 Married 2 Living Common Law 3 Never married (not living common law) 4 Separated (not living common law) 5 Divorced (not living common law) 6 Widowed (not living common law)	Variable name: **PDEM_11** 1 Married or living common law 2 Never married, separated, divorced or widowed (not living common law)
Immigration Status	Variable name: **IMMST** 1 Non-immigrant 2 Immigrant 3 Non-permanent resident	Variable name: **PIMMST** 1 Non-immigrant 2 Immigrant or non-permanent resident
Household living arrangements	Variable name: **HHLDARR** 1 Living alone 2 Multiple person household, no children 3 Multiple person household, with children 4 All members under 18 years old	Variable name: **PHHARR** 1 Living alone 2 Multiple person household, no children 3 Multiple person household, with children, with or without adults
Discrimination 2 years before COVID-19 Sex, Sexual orientation, Gender identity/expression	Variable name: **DIS_05_H** (Discrimination – Sex – 2 years before COVID-19) 1 Yes 2 No Variable name: **DIS_05_I** (Discrimination – Sexual orientation – 2 years before COVID-19) 1 Yes 2 No Variable name: **DIS_05_J** (Discrimination – Gender identity/expression – 2 years before COVID-19) 1 Yes	Variable name: **PDI05HIJ** (Discrimination – Sex, sexual orientation, gender identity/expression – 2 years before COVID-19) 1 Yes 2 No

	2 No	
Discrimination since COVID-19 began Sex, Sexual orientation, Gender identity/expression	Variable name: **DIS_10_H** (Discrimination – Sex – since COVID-19 bagan) 1 Yes 2 No Variable name: **DIS_10_I** (Discrimination – Sexual orientation – since COVID-19 began) 1 Yes 2 No Variable name: **DIS_10_J** (Discrimination – Gender identity/expression – since COVID-19 began) 1 Yes 2 No	Variable name: **PDI10HIJ** (Discrimination – Sex, sexual orientation, gender identity/expression – since COVID-19 began) 1 Yes 2 No

Appendix C

Renamed variables

Some variables were not recoded (i.e. they kept the same definitions and set of values in the Master and in the PUMF) but some of their values were suppressed. In these cases, variables were renamed on the PUMF.

Variable Description	Variable name on the Master File	Variable name on the PUMF
Discrimination because of indigenous identity – 2 years before COVID-19	DIS_05_A	PDIS_05A
Discrimination because of ethnicity/culture – 2 years before COVID-19	DIS_05_B	PDIS_05B
Discrimination because of race/skin colour – 2 years before COVID-19	DIS_05_C	PDIS_05C
Discrimination because of religion – 2 years before COVID-19	DIS_05_D	PDIS_05D
Discrimination because of physical/mental disability – 2 years before COVID-19	DIS_05_L	PDIS_05L
Did not experience discrimination – 2 years before COVID-19	DIS_05_N	PDIS_05N
Discrimination because of indigenous identity – Since COVID-19 began	DIS_10_A	PDIS_10A
Discrimination because of ethnicity/culture – Since COVID-19 began	DIS_10_B	PDIS_10B
Discrimination because of race/skin colour – Since COVID-19 began	DIS_10_C	PDIS_10C
Discrimination because of religion – Since COVID-19 began	DIS_10_D	PDIS_10D
Discrimination because of physical/mental disability – Since COVID-19 began	DIS_10_L	PDIS_10L
Did not experience discrimination – Since COVID-19 began	DIS_10_N	PDIS_10N
Difficulty - Seeing even when wearing glasses/contact lenses	LTC_05_A	PLTC_05A
Difficulty - Hearing even when using hearing aid/cochlear implant	LTC_05_B	PLTC_05B
Difficulty - Walking/stairs/using hands or fingers/other physical act.	LTC_05_C	PLTC_05C
Difficulty - Learning/remembering/concentrating	LTC_05_D	PLTC_05D
Difficulty - Emotional/psychological/mental health conditions	LTC_05_E	PLTC_05E
Difficulty - Other health problem/long-term condition	LTC_05_F	PLTC_05F
Difficulty - No health problem/long-term condition	LTC_05_G	PLTC_05G
Identifies as a person with a disability	LTC_10	PLTC_10
Disability status	DIS_STAT	PDISSTAT

Rural/Urban indicator	RURURB	PRURBAN
Indigenous identity flag	IIDFLAG	PIIDFLAG
Visible minority flag	VISMINFL	PVISMIN
Experienced discrimination in past 2 years	DISCRM2Y	PDIS2Y
Experience since COVID pandemic	DISCOVID	PDISCOV
Respondent is LGBTQ2	LGBTQ2	PLGBTQ2