Catalogue no. 89-657-X2025002
ISSN 2371-5006
ISBN 978-0-660-75072-9

Ethnicity, Language and Immigration Thematic Series

Survey on the Official Language Minority Population: User guide, 2022 (2025 revisions)

Release date: November 14, 2025

Statistique Canada
Statistics Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public".

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Table of contents

Survey on the Official Language Minority Population: User guide, 2022 (2025 revisions)

1.0 Survey description

The Survey on the Official Language Minority Population (SOLMP) was conducted by Statistics Canada in 2022 with the cooperation and support of Canadian Heritage. It is a postcensal survey of the English-speaking population in Quebec and the French-speaking population elsewhere in Canada. Questions were designed to assess changes in the official language minority populations since 2006, when a similar survey (Survey on the Vitality of Official-Language Minorities [SVOLM]) was conducted by Statistics Canada, and to provide new information on emerging issues regarding these minority populations.

The survey includes two broad samples: one for adults belonging to the official language minority population, and another composed of children who either have a parent belonging to an official language minority population or are themselves eligible for instruction in the minority official language.

Results of the survey were produced to support federal, provincial, territorial and municipal governments and community organizations in providing services, programs and initiatives for official language minorities. The data also serve researchers and other stakeholders interested in various aspects of French in Canada outside Quebec and English in Quebec.

Any questions about the dataset or its use should be directed to:

Statistics Canada
Centre for Demography, Language Statistics
Client Services
Email: statcan.languagestats-statlinguistique.statcan@statcan.gc.ca

2.0 Concepts and definitions

2.1 Survey on the Official Language Minority Population concepts and definitions

In this survey, the official language minority population refers to the English-speaking population in Quebec and the French-speaking population in Canada outside Quebec.

In this context, specific criteria were required to select the sample. These criteria, presented in detail in section 3.1 (Target and survey population), were determined in consultation with the external expert advisory committee put in place for the survey. They were selected to meet the following parameters: to allow for comparisons with the 2006 survey; to be sufficiently broad for users to specify subpopulations based on specific language criteria; and, for children, to account for new information on language of instruction. The criteria were not intended to create an official definition of the official language minority population.

The language concepts (mother tongue, language spoken most often at home, knowledge of official languages, eligibility for instruction in the minority official language) on which the criteria rely are all census concepts. Please refer to the census reference guides for definitions and quality notes related to these concepts:

- Languages Reference Guide, Census of Population, 2021
- Instruction in the Minority Official Language Reference Guide, Census of Population, 2021.

The SOLMP was designed to collect in-depth information on various aspects of official language minority populations to supplement information from existing data sources, including the census. In addition to sociodemographic information and household relationships, the survey covers the following main topics:

- language skills;
    - respondent's ability to understand, speak and write in English or French, as well as the hesitancy to use the official minority language, and perceived language insecurities and discrimination;

- education;
    - diplomas obtained and school attendance, including language of instruction, reasons for choosing a given language of instruction, and parents' intentions for the language of instruction of their children;
- early childhood services;
    - including language of daycare and reasons for choosing a given language for daycare;
- linguistic trajectory from childhood to adulthood and linguistic dynamics in a child's family;
    - languages spoken in early childhood and at age 15, and detailed information on languages spoken at home and languages spoken with friends;
- sense of belonging and perceived vitality;
    - sense of belonging to a language group, importance of English or French and their promotion for respondents, and perceived local presence of English or French when respondent was aged 15, currently (at the time of the survey) and over the past 10 years;
- civic participation;
    - participation in organizations, associations and networks within the community;
- use of language in the public sphere;
    - languages spoken outside the home and work, with people other than family or friends, in places such as grocery stores, restaurants and hospitals;
- government services;
    - language of communication with different levels of government (municipal, provincial or territorial, federal);
- contacts with the justice system;
    - language of interactions with the justice system;
- health services;
    - language of service in the health field;
- immigrant services;
    - language of services obtained when immigrating;
- art, culture and media;
    - language when using media and the Internet, participating in organized sports, attending arts events, etc.;
- geographic mobility;
    - place of birth, place of residence at 15 years of age and at the time of the survey, reasons and motives for geographic mobility;
- labour market participation;
    - detailed information on languages used at work, perceived obstacles at work related to languages, and job search services.

Questions related to the COVID-19 pandemic were added to the questionnaire, mostly to assess the extent to which the pandemic affected the survey results and might have contributed to changes since the 2006 SVOLM, particularly in the areas of English and French services and participation in activities.

SOLMP questions sometimes refer to the concepts of "community" and "municipality." These concepts are defined in the questionnaire as follows:

- "The term 'community' includes places that you visit regularly within a walking distance or short driving distance, such as the grocery store, schools, pharmacies or gyms."
- "The term 'municipality' refers to the city, the municipality or locality of residence as defined by provincial or territorial legislations."

The Data Dictionary along with the complete version of the User Guide will be accessible to microdata users, once the microdata files become available.

For detailed concepts associated with the SOLMP questions, please refer to the following documents:

- Surveys and statistical programs - Survey on the Official Language Minority Population (SOLMP)
- Questionnaire(s) and Reporting guide(s) – Survey on the Official Language Minority Population

2.2 Content development

The content of the 2022 SOLMP was based on the previous iteration of this postcensal survey, the SVOLM, conducted in 2006. Historical comparability was identified as a priority. While a large majority of the content was repeated, updates to the questionnaire were made based on consultations with Canadian Heritage and key data users, federal and community partners and external advisory committees.

These updates respond to new data needs and provide a better understanding of official language minority communities in Canada. Examples include a module on immigrants' experience with services to facilitate integration and new questions covering topics such as the hesitancy to use the official minority language, and linguistic insecurity and discrimination. Additionally, a new module on the impacts of the COVID-19 pandemic was created to assess the effect of the pandemic on activities and access to services in the minority official language.

The 2006 Census questionnaire was also reviewed with the aim of minimizing response burden. Approximately 80% of the questions and concepts from the 2006 questionnaire are included in the 2022 survey. For questions that were removed, some content can be derived, either partially or fully, through linkages with the census, thus reducing response burden.

Here are some examples of questions that were removed:

- education of parents in either official language, type of French program registered in, registration in an immersion program and number of years within a program;
- whether the child lives with both parents and, if not, the reasons and duration, and the frequency of stays in each house;
- grade of the child, registration date, mode of transportation to and from school, length of travel time, number of years in an immersion program, and reasons for changing programs;
- personal and household income.

Two rounds of qualitative testing was conducted in August 2020 and March 2021 with respondents who met the survey participation criteria. The objective was to ensure that all questions, particularly the new and modified questions identified as addressing new data needs, were easy to understand and could provide quality data. These tests were organized by experts in questionnaire design. The qualitative tests were held over a period of two weeks each, with approximately 30 interviews held in both English and French. Each interview lasted about one hour. Overall, the survey was well received by the respondents; questions were well understood, and, aside from some suggestions throughout, the feedback was positive.

While most of the modules underwent minor edits, a few required more significant changes following this testing. One major change was in the linguistic trajectory module, which was initially the last module of the questionnaire. Respondents felt the module was redundant when it was placed at the end, so it was moved to appear earlier in the questionnaire. The module itself was also modified. The geographic mobility module was perceived as convoluted and difficult to answer; therefore, major changes were made to simplify and shorten it. Finally, the education section was modified to facilitate understanding.

3.0 Survey methodology

Survey methodology and the analysis of complex survey data are broad and technical subjects, each deserving book-length treatment (which already exist). In Appendices C and E, the reader will find short, generic and non-technical overviews of both subjects intended as introductions to the more exhaustive expositions provided by textbooks and similar other sources. Appendix C provides a general methodological overview, which the following sections of this guide will expand upon with details specific to this survey.

The main takeaways of Appendix C are:

- The inaccuracy of an estimate admits two components: bias and variance.
- Unlike the variance, the bias is difficult (and usually is impossible) to assess in practice.
- For the SOLMP, proven methods and best practices were used to reduce bias-inducing factors in the data.
- The survey weight variable provided with the data allows users to reduce the bias in their estimates.
- The Bootstrap weight variables allow users to assess the variance associated with their estimates.
- The estimated variance is conveyed using a valid confidence interval.
- The narrower the valid confidence interval, the more accurate the estimate likely is.

3.1 Target and survey population

Since 1969, the *Official Languages Act* has recognized English and French as the country's **two official languages**. For people living in Quebec, English is the minority official language, while French is the minority official language for those living in any other province or territory.[1]

In the context of the SOLMP, a **person is considered part of the official language minority population** if they meet one of the following criteria:

1. their mother tongue includes the minority official language;
2. their mother tongue does not include English or French, and the only official language they know is the minority official language;
3. their mother tongue does not include English or French, they know both official languages, and they do not speak the majority official language most often at home.

Two main segments of the **population are covered** by the SOLMP: adults and children.

A person is included in **the adult component of the SOLMP** if they:

1. were 18 years of age or older on May 16, 2022 (the first day of SOLMP data collection);
2. reside in one of the 10 provinces or three territorial capitals, and do not live in a collective dwelling, on a First Nations reserve or in an Inuit community in northern Quebec;
3. are Canadian (non-permanent residents are excluded);
4. are part of the official language minority.

The **child component of the SOLMP** consists of two parts:

Children living with a parent who is part of the official language minority

In this part, children must meet the following criteria:

1. be younger than 18 years of age on May 16, 2022;
2. reside in one of the 10 provinces or three territorial capitals, and not live in a collective dwelling, on a First Nations reserve or in an Inuit community in northern Quebec;
3. not be a parent;

1. In this survey, the minority official languages of Canada are English in Quebec and French in Canada outside Quebec. However, in some provinces or territories, the languages with an official status may differ.

4. have at least one parent who meets the adult component requirements, except the parent must be 15 years of age or older.

It is worth noting that a child is considered to be part of the target population for the SOLMP based entirely on their parents' linguistic profile and not their own. Therefore, some surveyed children may not speak or understand the minority official language of the province or territory in which they live.

Children eligible for instruction in the minority official language

In this part, children must meet the following criteria:

1. be younger than 18 years of age on May 16, 2022;
2. reside in one of the 10 provinces or three territorial capitals, and not live in a collective dwelling, on a First Nations reserve or in an Inuit community in northern Quebec;
3. not be a parent.
4. Additionally, they must also meet one of the following criteria:

4a. they live outside Quebec within the same census family as a person aged 15 years or older who has French as their mother tongue;

4b. they live within the same census family as a person aged 15 years or older who is receiving or has received primary school instruction in the minority official language of their province or territory;

5. is receiving or has received instruction at the primary or secondary level in the minority official language of their province or territory;
6. live with a sibling within the same census family who is receiving or has received instruction at the primary or secondary level in the minority official language of their province or territory.

The concept of parenthood used here is broad. Beyond those who declared themselves a parent on the census, the SOLMP allows Canadians aged 15 years or older who live in the same census family to qualify as a parent. Thus, children eligible for instruction in the minority official language were purposely defined more broadly than they were for the 2021 Census. The main reason for this is because the specifications for the 2021 Census definition were not finalized at the time the SOLMP sample was drawn. A broader definition allowed for various scenarios to be covered by the SOLMP data, including the definition subsequently adopted by the census,[2] as well as other definitions of interest to users. The difference in population sizes between the broadest and the strictest definitions is estimated to be 1.7%.

In contrast, some children who are eligible for instruction in the minority official language by virtue of section 23 of the *Canadian Charter of Rights and Freedoms* and its jurisprudence could not be classified as such in the census and therefore could not be included in the SOLMP sample. While, by law, a child's eligibility for instruction in the minority official language depends on the characteristics of their parents and siblings, the census data establish family relationships only for people living in the same household. To know more about this limitation, consult Study on the Underclassification of Children Eligible for Instruction in the Minority Official Language in the 2021 Census.

Additionally, because of operational constraints in the territories, only the individuals residing in their capitals were eligible to be surveyed. According to the census, about 77% of French speakers in the territories reside in the capitals. Also, as per the requirements listed above, the SOLMP excludes non-Canadians and people living in a collective dwelling, on a First Nations reserve or in an Inuit community in northern Quebec.

2. Please refer to Instruction in the Minority Official Language Reference Guide, Census of Population, 2021 for more detail on the related census concepts.

3.2 Sample design

The SOLMP's sampling frame is derived from the data collected in the 2021 Census of Population, which was conducted on May 11, 2021, making the SOLMP a postcensal survey.

The 2021 Census uses two questionnaires[3] to collect data from people in a dwelling: the short-form questionnaire (Form 2A) and the long-form questionnaire (Form 2A-L or Form 2A-R depending on the geographic location of the dwelling). Form 2A-L is sent to approximately one in four private dwellings in most regions of Canada. In addition to the basic census demographic questions contained within the short-form (name, sex at birth and gender, date of birth and age, legal marital status, common-law status, relationship to Person 1, knowledge of official languages, languages spoken at home, first language learned, language of instruction, and Canadian military experience), Form 2A-L also includes questions on labour market activity, income, education, citizenship, housing, ethnic or cultural origins, religion, Indigenous identity, etc. Form 2A-R is similar to Form 2A-L and targets all dwellings in First Nations reserves, Métis Settlements, Inuit regions and other remote areas.[4]

Canadian citizens temporarily living abroad, full-time members of the Canadian Armed Forces stationed abroad, and visitors or representatives of foreign governments are excluded from the target population of the census long-form questionnaire. Also excluded are those living in collective dwellings (institutional or non-institutional). Collective dwellings include hospitals, residences for seniors, residential care facilities such as group homes for people with a disability or an addiction, shelters, correctional and custodial facilities, lodging and rooming houses, religious establishments, Hutterite colonies, establishments offering temporary accommodation services,[5] and other establishments.[6] For more information on collective dwellings, see the Dictionary, Census of Population, 2021.

While the census language questions—those of primary interest to the SOLMP—appear on the short-form and long-form questionnaires, the SOLMP's sample was selected from the people who answered the long-form questionnaire to enrich the SOLMP's answers with those obtained from the long-form. Exceptionally, wherever the number of respondents of the long-form questionnaire was insufficient to support the SOLMP's targeted analytical requirements, sampling was done from the larger pool of respondents of the short-form questionnaire.

The SOLMP's sampling frame was stratified by language region and age group. The frame underwent additional stratification, taking into account different factors for adults and children:

- For adults: The stratification was based on their language group.
- For children: The stratification was based on the language profile of their parents and whether they are eligible for instruction in the minority language.

As in 2006, the provinces of Quebec, Ontario and New Brunswick were split into six, five and three language regions, respectively, as detailed in the following table. For all other provinces, the language region corresponds to the province. For the three territories, the capital is each territory's sole language region. These are Whitehorse (CA) in Yukon, Yellowknife (CA) in the Northwest Territories and Iqaluit (CSD) in Nunavut.

3. See 2021 Census questionnaires.
4. In this document, the term "remote areas" is used to designate all these regions to lighten the text.
5. Such as a hotel, a campground, a YMCA or YWCA, a Ronald McDonald House, or an inn, when it is the respondent's usual place of residence on Census Day.
6. Like a student residence, military base, work camp or ship.

Table 1
SOLMP Regions and Corresponding Census Geography, for New Brunswick, Quebec and Ontario

Province	Census Metropolitan Area (CMA), Census Division (CD) or Census Subdivision (CSD)[1]	Region
New Brunswick	1312 (Victoria), 1313 (Madawaska), 1314 (Restigouche), 1315 (Gloucester), 1309036 (Alnwick), 1309038 (Neguac)	North
	1307 (Westmorland), 1308 (Kent), 1309016 (Rogersville, Parish), 1309017 (Rogersville, Village), 1309001 (Hardwicke)	Southeast
	All others not listed above	Rest
Quebec	462 (Montréal)	Montréal
	2401 (Îles-de-la-Madeleine), 2402 (Le Rocher-Percé), 2403 (La Côte-de-Gaspé), 2404 (La Haute-Gaspésie), 2405 (Bonaventure), 2406 (Avignon), 2495 (La Haute-Côte-Nord), 2496 (Manicouagan), 2497 (Sept-Rivières–Caniapiscau), 2498 (Minganie–Le Golfe-du-Saint-Laurent), 2407(La Matapédia), 2408 (La Matanie), 2409 (La Mitis), 2410 (Rimouski-Neigette), 2411(Les Basques), 2412(Rivière-du-Loup), 2413 (Témiscouata), 2414 (Kamouraska)	East
	2480 (Papineau), 2481 (Gatineau), 2482 (Les Collines-de-l'Outaouais), 2483 (La Vallée-de-la-Gatineau), 2484 (Pontiac), 2485 (Témiscamingue), 2486 (Rouyn-Noranda), 2487 (Abitibi-Ouest), 2488 (Abitibi), 2489 (La Vallée-de-l'Or)	West
	2430 (Le Granit), 2440 (Les Sources), 2441 (Le Haut-Saint-François), 2442 (Le Val-Saint-François), 2443 (Sherbrooke), 2444 (Coaticook), 2445 (Memphrémagog), 2446 (Brome-Missisquoi), 2447 (La Haute-Yamaska), 2448 (Acton), 2456 (Le Haut-Richelieu), 2468 (Les Jardins-de-Napierville), 2469 (Le Haut-Saint-Laurent)	Estrie and South
	2415 (Charlevoix-Est),2416 (Charlevoix), 2420 (L'Île-d'Orléans), 2421 (La Côte-de-Beaupré), 2422 (La Jacques-Cartier), 2423 (Québec), 2434 (Portneuf), 2417 (L'Islet), 2418 (Montmagny), 2419 (Bellechasse), 2425 (Lévis), 2426 (La Nouvelle-Beauce), 2427 (Robert-Cliche), 2428 (Les Etchemins), 2429 (Beauce-Sartigan), 2431 (Les Appalaches), 2433 (Lotbinière), 2491 (Le Domaine-du-Roy), 2492 (Maria-Chapdelaine),2493 (Lac-Saint-Jean-Est), 2494 (Le Saguenay-et-son-Fjord)	Québec and surrounding area
	All others not listed above	Rest
Ontario	3501 (Stormont, Dundas and Glengarry), 3502 (Prescott and Russell)	Southeast
	3506 (Ottawa)	Ottawa
	3548 (Nipissing), 3552 (Sudbury), 3553 (Greater Sudbury/ Grand Sudbury), 3554 (Timiskaming), 3556 (Cochrane), 3557 (Algoma)	Northeast
	3520 (Toronto)	Toronto
	All others not listed above	Rest

1. Consult 'Standard Geographical Classification (SGC) 2021 - Volume I, The Classification' for details.
Source: Standard Geographical Classification (SGC) 2021 - Volume I, The Classification.

The age groups for the provinces were 1 to 4 years, 5 to 11 years, 12 to 17 years, 18 to 24 years, 25 to 44 years, 45 to 64 years and 65 years or older, and the age groups for the capitals of the territories were 17 years or younger and 18 years or older.

A language group consists of people who are part of the official language minority, as defined above. However, in the Montréal area in Quebec, the basic language group was refined by adding the following two strata:

Individuals whose:

- Mother tongue is English and another non-official language, and whose known official language is only English; or,
- Mother tongue is English and another non-official language, and who know both official languages; or,
- Mother tongue includes English, French and another non-official language, and both official languages are known, and the language spoken most often at home is not French; or,
- Mother tongue does not include English or French, and the known official language is only English; or,
- Mother tongue does not include English or French, and both official languages are known, and the language spoken most often at home is English, but not French.

Individuals whose:

- Mother tongue does not include English or French, and both official languages are known, and the language spoken most often at home is a language other than the two official languages; or,
- Mother tongue includes English and French as well as another non-official language, and both official languages are known, and the languages spoken most often at home are the two official languages; or,
- Mother tongue includes English and French as well as another non-official language, and both official languages are known, and the language spoken most often at home is a language other than the two official languages.

It should be noted that in 2006, for specific analytical needs, a distinct sample of persons with a mother tongue other than French or English and having French as their first official language spoken was added in the Montreal region. In 2022, there was no such addition.

A simple random sample of people without replacement was selected in each stratum and independently across strata. For certain strata, notably those in Newfoundland and Labrador and Prince Edward Island, the targeted sample size exceeded the number of people available from the long-form questionnaire. In these cases, the sampling strategy was adjusted to include people from the short-form questionnaire. As a result, data from the long-form questionnaire will be available on the analytical file for only about one-quarter of the selected people in those strata.

Within a region, community experiences likely vary based on the concentration of the official language minority population, as measured by the proportion of people of the language minority living in each census dissemination area (DA). The sample design ensured that within a stratum, the sample represented well the full diversity of concentration levels by preventing an abundance of DAs with similar concentration levels.

3.3 Sample size

The sample size for a stratum was determined by targeting a level of accuracy comparable to that of the previous edition of this survey—the 2006 SVOLM—to estimate a proportion of about 8% with a coefficient of variation of 25% or, equivalently, with a margin of error of about 4 percentage points or with a confidence interval whose length is about 8 percentage points.

The following table provides sample sizes by language region and for the adult and children components of the SOLMP.

Table 2
Sample sizes for adult and children components and for whole survey, by province, territorial capital or region

Language region	Adults	Children	Region	Province or capital (territory)
Newfoundland and Labrador	949	964	1,913	1,913
Prince Edward Island	1,074	916	1,990	1,990
Nova Scotia	1,263	1,448	2,711	2,711
New Brunswick, north	1,297	971	2,268	6,910
New Brunswick, southeast	1,293	1,006	2,299	
New Brunswick, rest	1,236	1,107	2,343	
Quebec, east	1,110	1,026	2,136	20,209
Quebec, Estrie and South	1,281	1,343	2,624	
Quebec, Montréal	3,942	3,119	7,061	
Quebec, West	1,296	1,354	2,650	
Quebec, Québec and surrounding area	1,262	1,691	2,953	
Quebec, rest	1,264	1,521	2,785	
Ontario, northeast	1,297	1,163	2,460	12,827
Ontario, Ottawa	1,312	1,195	2,507	
Ontario, southeast	1,294	1,080	2,374	
Ontario, Toronto	1,295	1,417	2,712	
Ontario, rest	1,308	1,466	2,774	
Manitoba	1,269	1,299	2,568	2,568
Saskatchewan	1,166	1,426	2,592	2,592
Alberta	1,302	1,305	2,607	2,607
British Columbia	1,291	1,383	2,674	2,674
Whitehorse	332	391	723	723
Yellowknife	306	363	669	669
Iqaluit	251	203	454	454
Survey on the Official Language Minority Population	26,690	29,157	58,847	58,847

Source: 2022 Survey on the Official Language Minority Population.

4.0 Data collection

4.1 Collection period

In the 10 provinces, data collection took place from May 16 to December 16, 2022. In the three territorial capitals, the collection period was from August 22 to December 16, 2022.

As a postcensal survey, collection took place following the 2021 Census of Population, conducted on May 11, 2021. The Census allowed for the identification of the Official Language Minority population, who became the target population for SOLMP.

4.2 Collection instrument and modes of collection

As mentioned earlier, the survey questionnaire content was developed through consultations with, among others, members of an external advisory committee, as well as federal and community partners. The electronic questionnaire (EQ) was then developed with iterative rounds of testing. Two rounds of cognitive interviews were conducted in both official languages.

Once the SOLMP questionnaire was finalized, five methods were used to collect data:

- Respondent Electronic Questionnaire (rEQ). Respondents received a secure access code to log in and complete the survey online.
- Interviewer Electronic Questionnaire (iEQ), also known as Computer-Assisted Telephone Interview (CATI).
- As various restrictions were still in place related to the COVID-19 pandemic, Statistics Canada implemented CAPI Lite Plus collection (CLP). Interviewers visited selected individuals in person to schedule an appointment with them to later complete the questionnaire via CATI.
- '*Knock, Talk and Call*' (KTC), which is similar to CAPI Lite Plus, but a CATI interview is scheduled to take place immediately at the time of the visit.
- Finally, a pilot collection method was conducted for approximately one month in Iqaluit, where selected individuals were invited to complete their questionnaire via rEQ or CATI at a municipal building, where computers and in-person assistance were available.

Respondents could choose to complete the questionnaire in English or French. On average, the survey took about 45 minutes to complete.

4.2.1 Security of online survey questionnaires

The electronic data collection system for the 2022 SOLMP involved a secure web server for the Internet-based EQ questionnaire, which captured both the rEQ and iEQ survey data.

Statistics Canada takes very seriously the protection of confidential information provided by respondents, including online. A secure login process and robust encryption are key elements to prevent anyone from viewing or tampering with a respondent's survey information when it is completed and submitted online.

To protect the security of respondents' personal information when using the Internet, Statistics Canada incorporated the following safeguards:

- Strong **bi-directional encryption** technologies were used to ensure end-to-end security of data passing between the respondent's computer device and Statistics Canada's web server.
- Survey data were processed and stored on a restricted internal network which could only be accessed by those who had taken the oath of secrecy. Access is only possible on a "need-to-know" basis.
- Data submitted to Statistics Canada's web servers were encrypted before being stored and remained encrypted until they were transferred to the high security internal network.

Powerful firewalls, intrusion detection and stringent **access control** procedures were used to limit access to systems and databases.

4.3 Collection strategy

Those selected to participate in the SOLMP received an invitation letter in the mail describing the survey, informing them that they had been selected to participate and providing a secure access code for logging into the survey and completing it online. Each letter included a link to the SOLMP web page, as well a toll-free number to call if they wished to complete the survey by CATI or if they had questions (a teletypewriter (TTY) number was also provided for the hearing impaired). The letter also included a brochure, in English and French, which described the survey, the importance of participation and topics covered. In addition, a brochure in Inuktitut was included for residents of Iqaluit.

Reminders were sent by mail and text. Over the seven-month collection period, up to four reminder letters and six reminder e-mails were sent to respondents who had not yet submitted their questionnaire. Approximately 80% of respondents completed their survey using rEQ, with the remainder completing the survey using iEQ. Interviewers were instructed to make all reasonable attempts to obtain a completed interview with the selected respondent. Those who refused to participate were sent a letter to explain the importance of the survey and to encourage their participation and were re-contacted by telephone. Where possible, respondents in Iqaluit were visited in-person by an employee of Statistics Canada, who made contact, explained the survey and set up an interview appointment if necessary. Across Canada, respondents were interviewed in the official language of their choice, English or French.

4.4 Supervision and quality control

All Statistics Canada interviewers were under the supervision of senior interviewers who were responsible for ensuring that interviewers were familiar with the concepts and procedures of the survey to which they were assigned. Senior interviewers were also responsible for periodically monitoring the interviewers' work.

Interviewers were trained on the survey content, the CATI application, and on how to proceed with the new methods CLP and KTC. In addition to online or classroom training, the interviewers completed a series of mock interviews to become familiar with the survey and its concepts and definitions.

4.5 Proxy interviews

Proxy respondents (when someone fills out the questionnaire on behalf of a selected person) were not permitted for the SOLMP. However, for the child sample, parents were the targeted respondents, completing the questionnaire on behalf of their selected child.

4.6 Communications strategy

In the months leading up to data collection for the SOLMP, preparations were made to ensure that respondents had all the information they would need about the survey. A SOLMP survey webpage appeared on the Statistics Canada website. This webpage included background information on the survey and its methodology, and a link to the questionnaire. In addition, a special webpage of Information for Survey Participants, (ISP), was developed with step-by-step information on how to participate, and survey-specific "Questions and Answers".

A promotional campaign was developed. Community partners were also invited to encourage participation.

Promotional products that included handouts such as brochures and fact sheets were made available in English and French, and in Inuktitut for Iqaluit.

4.7 Special issues

4.7.1 Natural disasters

The SOLMP team closely monitored collection to be able to react proactively to specific situations, as needed. For example, following Hurricane Fiona which put several communities in Nova Scotia, New Brunswick and Prince Edward Island into local states of emergencies in September 2022, collection was paused in affected areas to reduce the burden on communities trying to recover. This involved turning off and on telephone priorities for CATI follow-up as communities were recovering.

4.7.2 COVID-19

The SOLMP was collected two years after the start of the COVID-19 pandemic, over approximately seven months. While impacts of the pandemic on data collection cannot be drawn precisely, this situation must be taken into consideration. Questions were added to the questionnaire to evaluate the impact of COVID-19 on certain questionnaire responses.

4.8 Response rates

The response rate was calculated by dividing the number of complete responses by the number of persons selected to participate minus the out-of-scope cases. Out-of-scope cases include people who, from when the 2021 Census was conducted to the time of the survey data collection, either died, emigrated, were institutionalized, or moved to a First Nations reserve. Also excluded were full-time members of the Canadian Forces living on a military base, visitors to Canada (misclassified during the census) or those who reported being less than 15 years of age at the time of completing the survey.

The response rate was 53.4% with 29,958 persons having participated. The table below provides the response rates (RR) and the respondent counts (#rep) by province, territorial capital or region, for the adult sample, the child sample, and for the whole survey.

Table 3
Response rates (RR) and respondent counts (#rep) for adult and children components and for whole survey, by province, territorial capital or region

	Adults		Children		Region		Prov./Cap. (territory)	
Region	RR	#rep.	RR	#rep.	RR	#rep.	RR	#rep.
Newfoundland and Labrador	41.7	345	50.6	484	46.5	829	46.5	829
Prince Edward Island	40.4	402	50.8	457	45.3	859	45.3	859
Nova Scotia	47.6	556	55.7	797	52.1	1,353	52.1	1,353
New Brunswick, North	44.0	556	49.8	481	46.5	1,037	51.4	3,447
New Brunswick, Southeast	51.8	644	57.8	580	54.5	1,224		
New Brunswick, Rest	49.8	570	56.4	616	53.0	1 186		
Quebec, East	53.0	567	45.7	455	49.5	1,022	58.5	11,219
Quebec, Estrie and South	65.3	802	56.5	745	60.8	1,547		
Quebec, Montréal	61.5	2,225	68.3	2,104	64.6	4,329		
Quebec, West	61.5	737	61.1	808	61.3	1,545		
Quebec, Québec and surrounding area	53.2	562	53.5	879	53.4	1,441		
Quebec, Rest	53.7	631	47.4	704	50.2	1,335		
Ontario, Northeast	48.6	603	52.5	608	50.5	1,211	52.8	6,464
Ontario, Ottawa	53.0	653	59.4	705	56.1	1,358		
Ontario, Southeast	49.8	625	56.2	604	52.8	1,229		
Ontario, Toronto	45.4	503	60.0	844	53.5	1,347		
Ontario, Rest	42.9	489	57.2	830	50.9	1,319		
Manitoba	50.5	598	55.2	711	52.9	1,309	52.9	1,309
Saskatchewan	46.7	491	53.8	758	50.8	1,249	50.8	1,249
Alberta	42.7	501	54.7	704	49.0	1,205	49.0	1,205
British Colombia	43.3	487	54.7	743	49.6	1,230	49.6	1,230
Whitehorse	48.5	146	53.4	206	51.2	352	51.2	352
Yellowknife	42.8	122	51.0	184	47.4	306	47.4	306
Iqaluit	32,8	77	29.4	59	31.2	136	31.2	136
Total SOLMP	**50.9**	**13,892**	**55.9**	**16,066**	**53.4**	**29,958**	**53.4**	**29,958**
French (Outside Quebec)	46.6	8,368	54.8	10,371	50.8	18,739	...	...
English (Quebec)	59.1	5,524	57.8	5,695	58.5	11,219	...	...
Atlantic	46.3	3,073	53.8	3,415	50.0	6,488	...	...
Quebec	59.1	5,524	57.8	5,695	58.5	11,219	...	...
Ontario	48.1	2,873	57.2	3,591	52.8	6,464	...	...
Prairies	46.6	1,590	54.5	2,173	50.9	3,763	...	...
British Colombia	43.3	487	54.7	743	49.6	1,230	...	...
Territorial Capitals	42.0	345	47.4	449	44.9	794	...	...
Provinces	51.1	13,547	56.2	15,617	52.0	29,164	...	...
Territorial Capitals	42.0	345	47.4	449	44.9	794	...	...

... not applicable

Source: 2022 Survey on the Official Language Minority Population.

5.0 Data processing

Processing transforms survey responses obtained during collection into a form that is suitable for tabulation and data analysis. It includes all data handling activities – automated and manual – after collection and prior to estimation.

5.1 Data capture

Responses to survey questions were captured in the electronic questionnaire (EQ), either directly by the respondent (rEQ) or by an interviewer (iEQ).

5.1.1 Respondent self-response (rEQ)

Respondents selected to participate in the SOLMP received a secure access code to log in and complete the survey online. Responses to survey questions were entered directly by the respondents into the electronic questionnaire (EQ). The responses were secured through industry standard encryption protocols, firewalls and encryption layers.

5.1.2 Interviewer-assisted response (iEQ)

Responses to survey questions were captured directly by the interviewer at the time of the telephone interview, using a computerized questionnaire which mimics the electronic questionnaire application. Similarly, the responses provided by respondents were secured through industry standard encryption protocols, firewalls and encryption layers.

For both the rEQ and iEQ, some editing was done directly at the time the EQ was being completed. Where the information was outside an acceptable range (too large or small) of expected values, or inconsistent with the previous entries, the respondent or interviewer was prompted, through message screens on the computer, to verify or clarify the information. However, for some questions the respondent or interviewer had the option of bypassing the edits and of skipping questions if the respondent did not know the answer or refused to answer. Consequently, the data were subjected to further edit processes after they were transmitted to head office. Once the electronic data were received, they were converted to readable text files.

5.2 Social survey processing steps

Data processing involves a series of steps to convert the electronic questionnaire responses from their initial raw format to a high-quality, user-friendly database involving a comprehensive set of variables for analysis. A series of data operations are executed to clean files of inadvertent errors, rigorously edit the data for consistency, code open-ended questions, create useful variables for data analysis, and finally to systematize and document the variables for ease of analytical usage.

The 2022 SOLMP used a set of social survey processing tools developed at Statistics Canada called the Social Survey Processing Environment (SSPE). The SSPE uses SAS software, custom applications and manual processes to perform the following systematic processing steps:

- Receipt of raw data
- Clean-up
- Recoding
- Coding
- Flows
- Edits and imputation
- Census linkage
- Derived variables
- Creation of final processing file
- Data certification
- Creation of dissemination files
- Creation of codebook (data dictionary)

5.3 Receipt of raw data to coding

Following the receipt of raw data from the SOLMP EQ application, a number of preliminary cleaning procedures were implemented at the individual record level. These included the removal of all personal identifier information from the files such as names and addresses, as part of a rigorous set of ongoing mechanisms for protecting the confidentiality of respondents. Duplicate records were resolved at this stage, and standard coding classifications were applied. Also, part of clean-up procedures was the review of all respondent records to ensure each respondent was “inscope” and had a sufficiently completed questionnaire. Note that the criteria to determine whether or not a respondent was in scope was applied before any edit or imputation was done.

5.4 Flows: response paths, valid skips and question non-response

Another set of data processing procedures for the survey was the verification of questionnaire flows or skip patterns. All response paths and skip patterns built into the questionnaire were verified to ensure that the universe or "target population" for each question was accurately captured during processing. Special attention was paid to distinctions between valid skips and nonresponse, which both read as a blank for that variable in the raw data. A valid skip is where a question was not relevant given a particular response to a previous question, and questionnaire flows were built so that the respondent was not asked that question. The respondent never saw it, so the question is coded as a valid skip. If a question was viewed and simply not answered, it is coded as not stated.

Special codes have been designated for each of these types of responses to facilitate user recognition and data analysis. The "valid skip" codes are set to "6" as the last digit, with any preceding digits set to "9" (for example, the code would be "996" for a three-digit variable). "Don't know" responses are coded to end in "7," with any preceding digits set to "9" (for example, "997"). "Not stated" codes end in 9, with any preceding digits also set to "9" (for example, "999"). These reserved codes are different for each variable, depending on how many categories the variable has and the length of the variable.

5.5 Coding open-ended questions

5.5.1 "Other-specify" items

Data processing also includes the coding of write-in responses, also referred to as "Otherspecify." Where a respondent's answer could not be easily assigned to an existing category, many questions allowed the respondent or interviewer to enter a writtenanswer text response in the "Otherspecify" category.

All questions with "Otherspecify" categories were closely examined during processing. Based on a qualitative review of the types of text responses given, coding guidelines were developed for each question. Based on these coding guidelines, many of the written answers provided were recoded back into one of the preexisting listed categories. Sometimes new categories may be created where there is sufficient numbers of responses to warrant them, but for the SOLMP that was never the case. Responses that were unique and different from existing categories were kept as "Other."

5.5.2 Open-ended questions and standard classifications

A few questions on the survey were asked in a completely openended format. These included questions related to the respondent's occupation and industry of work, where applicable. These responses were coded using a combination of automated and machine learning coding procedures. Standardized classification systems were used to code these responses.

5.6 Editing

Editing can occur at several points throughout the survey process and ranges from simple preliminary checks performed by the electronic questionnaire application to more complex automated verifications performed at the processing stage after the data have been collected. In general, edit rules are determined by what is logically or validly possible, based upon:

- expert knowledge of the subject matter;
- other related surveys or data;
- structure of the questionnaire and its questions; and
- statistical theory.

There are three main categories of edits: validity, consistency and distribution edits.

Validity edits verify the syntax of responses and include things such as checking that the data lie within an allowed range of values. For example, a range edit might be placed on the reported age of a respondent to ensure that it lies between 0 and 121 years.

Consistency edits verify that relationships between questions are respected. Consistency edits can be based on logical, legal, accounting or structural relationships between questions or parts of a question. The relationship between date of birth and marital status is one example where an edit may be applied: "a person less than 15 years of age cannot have any marital status other than single - never married."

Distribution edits are performed by looking at data across questionnaires. These edits attempt to identify records that are outliers with respect to the distribution of the data. Distribution edits are sometimes referred to as statistical edits or outlier detection.

5.7 Imputation

As detailed in Section 6, total nonresponse was addressed by reweighting the remaining respondents. No imputation was performed on account of partial nonresponse, to convert any missing value to a certain question into one of its admissible responses.

5.8 Linkage with Census (addition of Census variables)

Given that the SOLMP sample was drawn from the census database, respondent's responses to several relevant census questions and variables were added to the SOLMP database through data linkage. These variables are identified as 2021 Census of Population variables under the source line of a given variable in the data dictionary. Approximately 250 variables from the 2021 Census were linked to the final 2022 SOLMP analytical file. For all linked census variables, the census variable name was preserved as much as possible on the SOLMP database. However, some census variables were required to be renamed since the survey variable names are restricted to eight characters, whereas some census variable names exceed this limit. In these cases, there is a note in the data dictionary to indicate the original census variable name from which it was shortened.

5.9 Creation of derived variables

To facilitate data analysis, a number of data items on the microdata file were derived by combining items on the questionnaire. Answer categories of one, two or more variables may be combined to create a new variable. Derived variables may themselves be used as a basis for other derived variables.

For example, the variable "First official language spoken of the respondent" (PLOPRESP) was derived from answers to questions regarding the respondent's knowledge of official languages, mother tongue, and languages spoken most often at home.

The data dictionary identifies which variables were derived, as well as the primary variables on which these derived variables are based.

To facilitate more indepth analysis of the SOLMP dataset, approximately 200 derived variables were created by combining items on the questionnaire. Derived variables (DVs) were created across all major content domains.

Some simple DVs involved the collapsing of categories into broader categories. In other cases, two or more variables were combined to create a new or more complex variable that would be useful for data analysts. Some of the DVs were based on linked variables from the census, however if a respondent refuses census linkage, then their data are suppressed for census and census-based variables.

For most DVs, there is a residual category labelled "Not stated" for when the responses to the DV source questions do not meet the conditions to place a respondent in any of the valid categories for the DV. In many, but not all cases, a respondent is included in the "Not stated" category if any part of the equation was not answered (that is, if any question used in the DV had been coded to "Don't know" or "Not stated").

The 2022 SOLMP Data Dictionary identifies in detail which variables were derived and indicates the source variables from which the DVs were derived. A complete list of linked census variables and their accompanying notes are provided in the 2022 SOLMP Data Dictionary, which accompanies the survey analytical file.

5.10 Disclosure control

Statistics Canada is prohibited by law from releasing any information it collects which 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.

Specific instructions on how to release statistical information based on the SOLMP's data to meet these requirements are given in Section 8.

5.11 Creation of final data files and data dictionary (codebook)

Two final data files were created, one to cover the Adult component and the other to cover the Child component. The analytical files have been fully processed and prepared for release purposes.

A number of actions were taken to make final analytical files available to the public. Several measures were taken for the enhanced protection of respondent confidentiality. Personlevel weights were added to the file. Finally, all temporary variables or variables used exclusively for processing purposes were removed.

Accompanying the 2022 SOLMP analytical file is the record layout, SAS, SPSS and Stata syntax to load the file, and metadata in the form of a data dictionary that describes each variable and provides weighted and unweighted frequency counts.

The two analytical files were created for use in Research Data Centres (RDCs) and with the Real Time Remote Access (RTRA) service. The analytical files are available in RDCs across Canada but can be accessed only by researchers who fulfill certain requirements. Researchers associated with an academic institution, a government department or a nonprofit organization can also use the RTRA service to access the file. The analytical files can also be used at Statistics Canada to produce data tables in response to client customized requests.

Research Data Centres

The RDC program enables researchers to use the survey data in the master files in a secure environment at several universities across Canada. Researchers must submit research proposals that, once approved, give them access to an RDC.

Please see the following web page for more information: Research Data Centres

Real Time Remote Access

The RTRA system is an online remote access facility that allows users to run SAS programs, in real time, against microdatasets located in a central and secure location. Researchers submit SAS programs to extract results in the form of frequency tables. Confidentiality of the microdata is ensured in the RTRA system by deleting variables from the master file that do not contain a sufficient number of records to guarantee confidentiality. Researchers must complete an application form

Please see the following web page for more information: Real Time Remote Access

Custom tabulations

Another way to access the master files is to give all users the option of having SOLMP Client Services prepare custom tabulations. This cost-recovery product allows users who do not possess knowledge of tabulation software products to get custom results. The results are screened for confidentiality and reliability concerns before release.

For more information, please contact statcan.languagestats-statlinguistique.statcan@statcan.gc.ca.

6.0 Survey errors

Survey error is the numerical difference between the estimate obtained from the data and the unknown value of the population parameter of interest. Survey errors arise from a variety of sources, that can be classified into two main categories: sampling errors and non-sampling errors.

6.1 Non-sampling errors

Non-sampling errors are errors originating in any survey step other than sampling. Non-sampling errors are present in sample surveys and censuses (unlike sampling errors, which exist only in sample surveys). Non-sampling errors arise primarily from the following sources: non-response, coverage, measurement and processing. For example, a survey question whose meaning differs between French and English induces non-sampling error.

6.1.1 Non-response

Non-response occurs when complete and exact information could not be obtained from all selected individuals. Non-response not only reduces the amount of data for analysis but could also result in biased survey estimates if non-respondents and respondents have different characteristics.

Non-response is typically classified as total non-response or partial non-response. Total non-response occurs when information is missing or unusable for all[7] the survey variables, while partial non-response occurs when some of an individual's responses are missing or unusable.

Best practices were used before, during and after data collection to minimize the occurrence of non-response for the SOLMP. For example, before data collection began, the survey was promoted using brochures describing SOLMP and its benefits to targeted communities. During data collection, non-response was closely monitored to redirect efforts and resources where they were most needed.

For a given domain, the SOLMP's response rate is computed by dividing the number of respondents by the number of sampled individuals who are in-scope for the survey. The overall response rate was 53.4%, which is comparable with the rates obtained by other postcensal surveys conducted along a similar timeline. See the table in Section 4.8 for various other response rates of interest, and refer to the next section for a discussion of out-of-scope individuals.

Partial non-response may be addressed using a technique known as imputation, but it was not used for the SOLMP. Missing values to selected survey variables remain as "Not stated" in the analytical dataset. In social surveys such as the SOLMP, total non-response is typically mitigated through reweighting – see section 7.1 for further details.

6.1.2 Coverage errors

Coverage errors consist of omissions, erroneous inclusions, duplications and misclassifications of units in the survey frame. Coverage errors may cause a bias in the estimates and the effect can vary for different sub-groups of the population.

7. A person may have provided valid data to some questions—such as those in the survey's identification module—yet still be considered a non-respondent. A respondent is a person who has provided valid information to a pre-identified set of core survey variables known as "the survey's minimal content."

Some adults initially considered to be of interest according to census data fell out of the survey's scope for a variety of reasons based on their answers to the SOLMP, including 1) they were no longer part of the official language minority on May 16, 2022 (a year after the census), 2) they moved outside Canada, 3) they were institutionalized or 4) they were deceased. For children, the same factors apply except for the first one, which was operationally too difficult to assess, as a child's eligibility to be surveyed is based on the language profile of their parents and not just on the person responding to the SOLMP on their behalf.

Out-of-scope rates expressed in percentage points—calculated as the number of out-of-scope individuals divided by the number of sampled individuals with whom a contact was established during collection—are provided in the table below.

For adults, the change in official language minority status is responsible for over 80% of all out-of-scope cases (a figure not shown in the table). These are mostly individuals recorded by the census as speaking both official languages, but who reported not speaking the minority language in the SOLMP. This may be attributed in part to the proxy effect whereby one person answers the census questionnaire on the behalf of all individuals in their household, not fully knowing everyone's language profile.

Table 6.1
Out of Scope Rates: Overall and excluding adults whose Official Language Minority status changed

Region	Adults	Children	Region[1]	Prov./Cap. (territory)[1]
	percent			
Newfoundland and Labrador	16.4	1.0	8.4 (1.2)	8.4 (1.2)
Prince Edward Island	9.3	2.1	5.9 (1.4)	5.9 (1.4)
Nova Scotia	9.5	1.5	5.1 (1.3)	5.1 (1.3)
New Brunswick, North	3.6	0.8	2.3 (1.2)	3.7 (1.2)
New Brunswick, Southeast	4.9	0.4	2.9 (1.1)	
New Brunswick, Other	9.7	1.5	5.7 (1.4)	
Quebec, East	4.3	3.6	4.0 (2.5)	5.7 (2.0)
Quebec, Estrie and South	4.7	2.2	3.4 (1.6)	
Quebec, Montréal	9.2	1.3	5.7 (1.8)	
Quebec, West	8.5	2.7	5.5 (2.0)	
Québec CMA of Québec	18.7	3.2	9.8 (2.7)	
Quebec, Rest	8.2	2.9	5.4 (1.9)	
Ontario, Northeast	5.7	0.7	3.2 (1.1)	5.7 (1.1)
Ontario, Ottawa	7.9	0.8	4.4 (0.8)	
Ontario, Southeast	4.0	0.7	2.5 (0.9)	
Ontario, Toronto	19.1	0.9	9.1 (1.0)	
Ontario, Rest	17.2	1.3	8.5 (1.5)	
Manitoba	8.9	1.0	4.7 (1.0)	4.7 (1.0)
Saskatchewan	13.0	1.5	6.4 (1.5)	6.4 (1.5)
Alberta	14.0	1.6	7.4 (1.6)	7.4 (1.6)
British Columbia	17.5	2.2	9.2 (1.9)	9.2 (1.9)
Whitehorse	10.4	1.4	5.5 (1.9)	5.5 (1.9)
Yellowknife	7.8	0.6	3.8 (1.0)	3.8 (1.0)
Iqaluit	7.5	1.1	4.6 (0.8)	4.6 (0.8)
SOLMP	**10.0**	**1.6**	**5.7 (1.5)**	**5.7 (1.5)**

1. At the region and the province-or-territorial-capital levels, two out-of-scope rates are reported: the overall rate and, in parentheses, the rate obtained by excluding a change in the official language minority status reported by certain adults—the first factor mentioned in Section 6.1.2. From these rates, one can infer that out-of-scope rates for adults are comparable with those of children once that factor is excluded from the analysis.

Source: 2022 Survey on the Official Language Minority Population.

As mentioned in Section 3.2, for sampling purposes children were classified in one of several groups based on the language and education profile of their parents. For operational reasons, SOLMP's sampling was conducted using a preliminary classification based on the collected census data. Later when the classification was complete, some children were reclassified and the survey weight adjusted accordingly.

6.1.3 Measurement errors

Measurement errors (sometimes called response errors) occur when the response provided differs from the answer the respondent should have given. Such errors may be attributable to the respondent, the interviewer, the questionnaire, the collection method or the respondent's record-keeping system. These errors may be random, or if not, they may introduce systematic bias.

Questions selected for potential inclusion in the 2022 SOLMP questionnaire underwent several rounds of qualitative testing using one-on-one virtual interviews with respondents in various communities across Canada. Qualitative testing of the survey questionnaire was carried out by Statistics Canada's Questionnaire Design Resource Centre (QDRC). Based on the testing results, adjustments were made to the wording and the sequence of the questions to reduce measurement error.

The SOLMP questionnaire also used harmonized content—commonly asked questions formatted in a standard way for use across all surveys.

Many other measures were taken to specifically reduce measurement error, including the use of skilled interviewers, extensive training of interviewers in survey procedures and content, as well as observation and monitoring of interviewers to detect problems of questionnaire design or misunderstanding of instructions. In addition, incoming data were evaluated in real time to detect problems of questionnaire design or misunderstanding of instructions.

It is costly to accurately measure the level of response error, and few surveys conduct a post-survey evaluation. However, interviewer feedback and observation reports usually provide clues as to which questions may be problematic (poorly worded question, inadequate interviewer training, poor translation, technical jargon, lack of available help text, etc.).

6.1.4 Processing errors

Processing error is the error associated with activities conducted once survey responses have been received. It includes all data handling activities after collection and before estimation. Like all other errors, they can be random, inflating the variance of the survey's estimates, or systematic, introducing bias. It is difficult to obtain direct measures of processing errors and their effect on data quality, especially as they are mixed with other types of errors (non-response, measurement and coverage).

Processing errors may occur at various stages of the survey process, including programming the electronic questionnaire, data capture by the interviewer or respondent, coding, and data editing. Quality control procedures were applied to every stage of the SOLMP data processing to minimize this type of error. Data were collected using an electronic questionnaire, either administered by an interviewer or completed by self-reporting over the Internet. Several verifications were built into the system to warn the respondent or the interviewer in the event of inconsistencies or unusual values, making it possible to correct them immediately (see Section 5.6).

At the data processing stage, a detailed set of procedures and edit rules were used to identify and correct any inconsistencies between the responses provided. For each step of data cleaning, a set of thorough, systematized procedures was developed to assess the quality of every variable in the file and correct any errors found. A snapshot of the output files was taken at each step and verification was done by comparing files at the current and previous steps. The programming of all edit rules was tested before being applied to the data. Examples of data processing verifications included:

- the review of all question flows, including very complex sequences, to ensure skip values were accurately assigned and distinguished from different types of missing values
- an in-depth qualitative review of open-ended and "Other — Specify" responses to ensure accurate and rigorous coding
- the completion of coding to standardized classifications
- a review of question level non-response rates

- the review of all derived variables against their component variables to ensure accurate programming of derivation logic, including very complex derivations.
- When possible, a comparison to the 2006 survey iteration and to other data sources was also performed.

See the data processing section of this guide for further details (Section 5).

6.2 Sampling errors

Sampling error is survey error that comes from estimating a population parameter of interest based on data from the sampled individuals, which represent only a portion of all individuals of interest. While sampling error can manifest itself as bias and variance, best practices and proven methods are used to keep the bias to a minimum, leaving only the variance to be assessed. For the SOLMP, a bootstrap method adapted to its key design features was used to estimate the variance.

Factors affecting the magnitude of the variance include the following:

1. The spread of the characteristic of interest in the population: The more spread out the characteristic in the population, the larger the variance.
2. The size of the population: In general, population size affects only the variance of small to moderately sized populations.
3. The sample size: All other things being equal, the variance decreases as the sample size increases. (This provides a convenient way to reduce some of the survey's total error.)
4. The sample design and method of estimation: Some sample designs are more efficient than others in the sense that, for the same sample size and method of estimation, one design can lead to smaller variance than another. A less efficient sample design may be used in situations where it is operationally more convenient.

While the variance measures all the sampling errors in bias-free estimates, it is not directly used when reporting on their accuracy, as the standard error, the coefficient of variation (CV) and the confidence interval (CI)—all by-products of the variance—are more convenient and informative.

The standard error[8] of an estimator is the square root of its variance. This measure is easier to interpret than the variance because it is on the same scale as the estimate itself. For example, the standard error of an estimate of income expressed in dollars will also be in dollars, not in "dollars squared" as with the variance. However, the standard error does not lend itself well to comparing the accuracy of two estimates. For example, when the size N of the population is known, dividing the estimate of a total of interest by 1 (which simply yields the estimated total) or by N (which turns the estimated total into an estimated mean) does not affect its accuracy because both are constants. However, the standard error of the estimate of the mean is N times smaller than that of the corresponding estimated total, suggesting instead that the former is more accurate than the latter.

The CV of an estimate is a relative measure of the error. It is calculated as the standard error divided by the estimate of interest. It is dimensionless and is usually expressed as a percentage (e.g., 10% instead of 0.1). Returning to the previous example, one can verify that the CV of the estimated proportion (with a constant denominator) is equal to the CV of the estimate for the count of interest. While the CV is useful for comparing the sampling errors of large positive estimates, it is not recommended for estimates of proportions, change, differences, or estimates close to zero or in negative values.

At Statistics Canada, reporting the sampling error of an estimate using a 95% CI is considered a best practice. The 95% CI of an estimate means that if the survey were repeated, 95% of the CIs obtained (or 19 times out of 20, to use the well-known expression from voting intention polls) would contain the true population value.

8. This term is often used interchangeably with "standard deviation", and while the two are related they are not equivalent. The standard deviation is the square root of some variance. The standard error of an estimator is the standard deviation of its variance.

This is the usual probabilistic interpretation of a CI. However, the reader may find the following heuristic argument more informative: Because of survey error, one cannot say with certainty whether the estimate obtained from the data is close to the true value. In other words, while the estimate represents one plausible value for the parameter of interest based on the data, there is a whole range of other plausible values around the estimate because of survey error. This range of plausible values is the 95% CI. The shorter the 95% CI, the narrower the range of plausible values (including the estimate) for the parameter of interest, and the stronger the likelihood that the estimate is close to the true value .

Appendix D provides general information about confidence intervals in the context of postcensal surveys, while Section 8.6 provides more specific instructions on how to obtain a valid 95% CI from the SOLMP data.

7.0 Survey weights and Bootstrap weights

The principle behind estimation in a probability sample is that each unit selected in the sample represents, besides itself, other units that were not selected in the sample. For example, if a simple random sample of size 100 is selected from a population of size 5,000, then each unit in the sample represents 50 units in the population (including itself). The number of units represented by a unit in the sample is called the survey weight of the sampled unit.

The following section provides the details of the method used by the survey team to calculate the survey weight variable and the bootstrap weight variables for the SOLMP. The survey weight variable is used to mitigate the bias that surveying activities (e.g., the sample design, non-response) may be introducing in the estimates. To get an indication of the reliability of a weighted estimate, 1,000 bootstrap weight variables are supplied to assess the variance that is to be reported in the form of a CI. Section 8.6 provides information on using the bootstrap weights.

7.1 Survey weight variable

Because the SOLMP is a postcensal survey, its sample design consists of several phases starting with a stratified sample drawn from the frame created using the answers collected from the census. Thus, an initial sampling weight variable is derived from the census weight variable and the SOLMP's sampling fractions.

For various operational reasons, some sampled individuals are removed from data collection activities, including in situations where

1. more than three people in a same household have been selected for the SOLMP
2. more than three people in a same household have been selected when considering the other two postcensal surveys, the Indigenous Peoples Survey (IPS) and the Canadian Survey on Disability (CSD)
3. samples from the SOLMP and IPS overlap
4. available contact information is insufficient to support data collection
5. multiple records linked to the same person have been identified by analyzing names, dates of birth and addresses.

In the first two cases, affected sampled individuals were randomly removed, and their sampling weight redistributed among those remaining. In the other cases, the concerned individuals (or excess records) were removed and their sampling weight redistributed among those remaining.

The reweighting performed for the SOLMP to address total non-response assumes that non-response, assimilated to a random mechanism, can be explained solely by using the information available to the survey team about respondents and non-respondents, and is therefore not dependent on any unmeasured factors. This explains why, whenever feasible, the SOLMP's sampling was carried out using respondents of the census long-form questionnaire rather than the short-form questionnaire, to take advantage of the detailed sociodemographic information available on the long-form questionnaire about the SOLMP's respondents and non-respondents.

From an operational standpoint, three cases were handled separately and in succession: 1) individuals not contacted, 2) total non-response after contact, and 3) partial non-response after contact. The first case consists

of sampled individuals who could not be contacted despite the availability of their census contact information and the survey team's best efforts. Their weight was redistributed among those who were successfully contacted. Total non-response comes from sampled individuals who were successfully contacted but failed to provide the basic required survey information. Their weight was redistributed among those who responded. Finally, partial non-respondents are individuals who did not provide the necessary information to be respondents. These cases were converted to total non-response.

The non-response adjustment to the weights was done in three steps. In the first step, a logistic regression model was used to predict the response probability (probability of getting a response) for each unit (both responding and non-responding units) from a series of explanatory variables. Explanatory variables are "person" or "household" characteristics from the 2021 Census (e.g., the number of people in the selected person's household) for the selected person or the person most knowledgeable in the case of children.

In the second step, respondents and non-respondents with similar predicted response probabilities were grouped into adjustment classes using a classification analysis. The response rate was derived for each class based on the number of respondents and non-respondents in the class. The calculated response rate was then weighted using the weights from the previous adjustment step.

In the third and final step, the weights of the responding units within each class were adjusted using the inverse of the weighted response rate in that class. The weights of the non-responding units were set to zero.

At the final stage of the creation of the survey weight variable, post-stratification was performed by the region used for stratification purposes, age group and language group based on the respondents' updated classification information to align corresponding estimates with those of the census. The resulting survey weight variable was examined to identify extreme values that needed adjustment.

The SOLMP's weight variable is called WEIGHT.

The SOLMP uses a complex sample design, and this translates into unequal survey weights between individuals in the analytical file. When producing estimates and statistical tables, analysts must make proper use of the survey weight variable to draw statistically meaningful conclusions from the SOLMP's data. Failing to use the survey weight variable may lead to biased estimates and will be inconsistent with those released by Statistics Canada.

7.2 Bootstrap weight variables

The exact variance cannot be computed because it would require an estimate from each admissible sample when, in practice, data are collected only for one such sample, the sample randomly selected by the survey. Still, that variance can be estimated using the bootstrap weight variables supplied with the analytical file.

A bootstrap procedure adapted to the SOLMP's design produced 1,000 bootstrap weights, bsw1-bsw1000, which are to be used to assess the variance associated with the estimate. In principle, analysts obtain these 1,000 bootstrap estimates by replacing in turn the survey weight variable with each of the supplied 1,000 bootstrap weight variables in their statistical analysis. In practice, analysts use a suitable software to get a bootstrap variance estimate from the data and the bootstrap weights provided – see Section 8.6.1 for details.

Important:

- Even users with extensive expertise of bootstrap methodology must use the SOLMP's supplied set of 1,000 bootstrap weight variables, rather than undertaking a bootstrap implementation of their own. This is because the analytical file does not contain all the survey design information that these users would need to implement the bootstrap to adequately capture the contribution of the various components of the SOLMP to the variance.

- Because of the specific bootstrap procedure used for the SOLMP, an expansion multiplicative factor of 16 must be used for the computed bootstrap variance to properly estimate the exact variance. See Section 8.6 for more details.
- Despite what may be inferred from the software documentation, simply specifying the SOLMP's stratified design is not enough to ensure that the computed variance estimates are appropriate; the bootstrap weight variables must be used. Therefore, it is not enough to specify the survey weight variable into the WEIGHT statement of the software and let it perform variance estimation without using the supplied bootstrap weight variables. See Section 8.6 for more details.

8.0 Guidelines for tabulation, analysis and release

This chapter of the documentation outlines the guidelines to be adhered to by users tabulating, analyzing, publishing or otherwise releasing any data derived from the survey microdata files. With the aid of these guidelines, users of microdata should be able to produce the same figures as those produced by Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

Appendix E provides an overview of how to analyze complex survey data.

8.1 Confidentiality guidelines

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.

Confidentiality vetting rules are applied to all Statistics Canada survey results before the results are made public, regardless of the mode of data access. They apply equally to Statistics Canada employees and deemed employees in the Research Data Centres (RDCs). These rules are designed to ensure the confidentiality of respondents.

For researchers in the RDCs, there are two levels of confidentiality screening for SOLMP estimates prior to dissemination. First, any researcher accessing the confidential analytical files must follow the dissemination guidelines outlined in this User Guide. Secondly, all results will be vetted by a designated RDC analyst. **Note that supplementary vetting rules need to be applied to all income variables.** Whether or not results are removed from an RDC depends on whether the output meets the guidelines on confidentiality. Once dissemination and release guidelines provided in the survey documentation have been followed, decisions about whether results should be reported in a publication are up to the discretion and responsibility of the researchers and the peer review process.

Table 8.1 below summarizes the confidentiality guidelines for the SOLMP. All data must be released in aggregate form. In general, the release of unweighted data is prohibited. Unweighted frequencies underlying weighted estimates must be 10 or higher. Rounding is required for all weighted estimates. **For estimates pertaining to 2021 Census income or estimates pertaining to detailed geographies that are below the level of a SOLMP region, more restrictive rules apply**. Section 8.2 provides more information concerning minimum sample sizes.

Table 8.1
Confidentiality guidelines for the 2022 Survey on the Official Language Minority Population

Criterion	2022 SOLMP Guideline	Notes
1) What is the minimum required unweighted frequency?	10	For income estimates, there must be at least 10 units with non-zero income. Also see section 8.2 below.
2) Is unweighted descriptive output allowed?	NO	Unweighted descriptive outputs (e.g., totals, proportions, ratios, means, correlations) are prohibited. Model N's and sample sizes may be released with justification from the principal investigator. Also see (3) below.
3) May unweighted and weighted descriptives both be released for this survey?	NO	Permission will usually be given ONLY in the case in which a journal requires both weighted and unweighted frequency tables for publication (letter from editor required).
4) May both unweighted and weighted model output be released for this survey?	YES	…
5) What is the definition of "Detailed Geography" for this survey?	Any tabulation below the language region is considered "detailed". The Census Metropolitan Areas (CMAs) of Toronto, Montreal and Vancouver are exceptions – tabulations for other CMAs and other sub-provincial/ territorial geographical levels must conform to the requirement.	For a justification of the underlying concept, see RDC for details. Also see sections 8.2 and 8.3 below.
6) Restricted geographical variables	Sampling frame information (e.g., lists of postal codes or Census SubDivisions (CSD)) below the detailed level of geography for a survey is considered confidential and may not be released.	This criterion refers to surveys with a restricted level of geography, below which it is possible to identify survey frame information. It is included in every survey's documentation, but it does not apply to the 2022 SOLMP. See RDC for the Disclosure Analysis Abridged Rules Section VII.
7) Is rounding required for all weighted descriptive outputs?	YES	100 for income totals
If yes, what is the rounding base?	Rounding base ·10 for counts and totals ·100 for income totals ·10 for income means or medians ·1 (i.e., to the nearest integer) for ratios involving income variables. The number of decimal places for proportions is restricted to three (0.001 or 0.1% for percentages).	·50 for counts and totals ·100 for income totals ·10 for income means or medians ·1 (i.e., to the nearest integer) for ratios involving income variables. The number of decimal places for proportions at detailed levels of geography is restricted to two (0.01 or 1% for percentages). Also see section 8.3 below.
8) What confidentiality rules should be applied when producing estimates for 2021 Census income as part of the 2022 SOLMP?	RDCs should refer to document on confidentiality rules for the 2021 Census which should be applied in addition to the confidentiality rules of the 2022 SOLMP. Client Services should refer to the income rule documented internally under separate cover in addition to the confidentiality rules noted in this guide.	Also see sections 8.2 and 8.3 below.

... not applicable

Note: For definitions of CMA and SOLMP geographies, see the Census Dictionary.

Source: 2022 Survey on the Official Language Minority Population.

8.2 Minimum sample size guidelines

For the SOLMP, a minimum sample size must be respected to meet the confidentiality requirements of the *Statistics Act*. This minimum sample size is also important for the reliability of estimates. For the 2022 SOLMP, the following minimum applies for unweighted frequencies for all descriptive statistics.

- The minimum unweighted frequency count must be 10.
- In any given frequency distribution, only cells respecting the minimum unweighted frequency criterion can be disseminated.
- All other types of descriptive statistics must be calculated from at least this minimum number of observations. If the descriptive statistic is bivariate, such as a ratio or a difference, then both contributing variables must have at least this minimum number of observations to contribute. For example, if a ratio is produced, both the numerator and the denominator must be based on at least the minimum number of observations.
- For income estimates (e.g., medians, averages, totals), there must be at least 10 units with non-zero income in a given cell for it to be published.

8.3 Rounding guidelines

To ensure that estimates produced from SOLMP analytical files will correspond to those produced by Statistics Canada, the user is strongly advised to follow the rounding guidelines provided below. Disseminating unrounded estimates could be misleading, since such estimates might appear to be more precise than they are. For confidentiality purposes, rounding is required for all weighted descriptive statistics.

1. Estimates of totals that appear in the body of a statistical table should be rounded to the nearest ten by the traditional rounding method (see description of method below).
2. Partial and grand totals in statistical tables should be calculated from their unrounded components, then rounded to the nearest ten by the traditional rounding method.
3. Averages, proportions, rates, percentages and other forms of ratios should be calculated from unrounded components, but then be rounded to one decimal place by the traditional rounding method. For example, a value of 0.1234 or 12.34% should be rounded to 0.123 or 12.3%. If working with detailed geography (see Table 8.1), averages, proportions, rates, percentages and other forms of ratios should be rounded to the nearest integer by the traditional rounding method. For example, a value of 0.1234 or 12.34% should be rounded to 0.12 or 12%.
4. Sums and differences of aggregates or ratios should be calculated from their corresponding unrounded components, then rounded to the nearest ten or the nearest decimal using the traditional rounding method.
5. Regardless of geography, weighted income totals should be rounded to the nearest 100 and weighted income means or medians should be rounded to the nearest 10. Weighted ratios involving income variables should be rounded to the nearest integer.
6. Confidence intervals for estimates should be calculated from their unrounded components, and then rounded to the nearest ten or to one decimal place by the traditional rounding method. (Since the estimate and the corresponding confidence limits are rounded independently, the estimate will not always appear exactly in the middle of the confidence interval.)
7. For detailed geographies, all statistics must be rounded to the nearest 50. For the 2022 SOLMP, any tabulation below the language region is considered “detailed”. The CMAs of Toronto, Montreal and Vancouver are exceptions - tabulations for other CMAs and other sub-provincial/territorial geographical levels must conform to the requirement. At this detailed level, remember to round all forms of ratios to the integer (see point 3 above).
8. In the event of technical or other constraints, a rounding method other than traditional rounding may be used. In such cases, the estimates obtained may differ from the corresponding estimates produced by Statistics Canada. If so, the user is strongly advised to state the reason for these differences in the document disseminated.

8.3.1 The traditional rounding method

According to the traditional rounding method, if the first or only digit to be suppressed falls between 0 and 4 (e.g. the "3" in "823" when rounding to the nearest 10 or the "2" when rounding to the nearest 100), the last digit retained does not change (e.g. the "2" in "823" remains the same when rounding to the nearest 10, resulting in "820" or the "8" remains the same when rounding to the nearest 100, resulting in "800"). If the first or only digit to be suppressed falls between 5 and 9 (e.g. the "5" in "865" when rounding to the nearest 10 or the "6" when rounding to the nearest 100), the value of the last digit retained is increased by one unit (1) (e.g. the "6" in "865" is increased by one unit when rounding to the nearest 10, resulting in "870" or the "8" is increased by one unit when rounding to the nearest 100, resulting in "900").

8.4 Sample weighting guidelines for tabulation

The SOLMP uses a complex sample design and estimation method, and the survey weights are therefore not equal for all the sampled units. When producing estimates and statistical tables, users must apply the survey weights. If the survey weights are not properly used, the estimates derived from the microdata files cannot be considered representative of the survey population and will not be consistent with those produced by Statistics Canada.

All respondents whose data appear in the analytical file met a pre-established survey response standard. However, some missing data remain because of non-response for some questionnaire items (i.e., partial non-response). These are situations where respondents did not answer a particular question that applied to them.

The "Valid skip" category is generally not considered a missing value since it indicates a question that was not intended for that particular respondent. The same is true for the "Not applicable" categories for census variables, which are equivalent to valid skips. The term "Valid skip" is not used in the census and, consequently, census variables in the SOLMP analytical file maintain the same category labels they were given for the census.

Including or excluding any of the aforementioned types of missing values in a tabulation depends on the analysis objective. Users will need to define their estimation domain (total population of interest) for each variable, considering the missing values for each. For example, they will have to assess the relevance of including missing values in the denominator used to calculate percentages. In some cases, researchers may decide that missing values are meaningful for their research question. The fact that a respondent chooses not to answer a question may constitute in itself useful information.

8.5 Release guidelines for quality

Before releasing and/or publishing any estimates, analysts should consider the quality level of the estimate. While data quality is affected by both sampling and non-sampling errors, this section covers quality in terms of sampling error. It is considered a best practice at Statistics Canada to report the sampling error of an estimate through its 95% confidence interval (CI). The confidence interval should be released with the estimate, in the same table as the estimate. In addition to the confidence intervals, estimates are categorized into one of three release categories:

Category A

The estimate and confidence interval can be released with no warning. Data users should use the 95% confidence interval to assess whether the quality of the estimate is sufficient. Note that the 'A' is not a quality indicator; it should not be released with the estimate.

Category E

The estimate and confidence interval should be flagged with the letter E (or some similar identifier) and accompanied by a quality warning, indicating that estimates should be used with caution. Data users should use the 95% confidence interval to assess whether the quality of the estimate is sufficient.

Category F

The estimate and confidence interval are not recommended for release. They are deemed of such poor quality, that they are not fit for any use; they contain a very high level of instability, making them unreliable and potentially misleading. If analysts insist on releasing estimates of poor quality, even after being advised of their accuracy, the estimates should be accompanied by a disclaimer. Analysts should acknowledge the warnings given and undertake not to disseminate, present or report the estimates, directly or indirectly, without this disclaimer. The estimates should be flagged with the letter F (or some similar identifier) and the following warning should accompany the estimates and confidence intervals: “Please be warned that these estimates and confidence intervals [flagged with the letter F] do not meet Statistics Canada’s quality standards. Conclusions based on these data will be unreliable, and may be invalid.”

The rules for assigning an estimate to a release category depends on the type of estimate.

8.5.1 Release Rules for Estimates

Estimated proportions and estimated counts are computed from binary variables. Estimated counts are estimates of the total number of persons/households with a characteristic of interest; in other words, they are the weighted sum of a binary variable (e.g., estimated number of adults in the official language minority). Estimated proportions are estimates of the proportion of persons/households with a characteristic of interest (e.g., estimated proportion of adults in the official language minority population with respect to the Canadian population). Estimated counts and proportions can also be computed from categorical variables: that is, estimates of the number or proportion of persons/household who belong to a category.

The release rules for estimated proportions are based on sample size and the length of the confidence interval. The release rules for other statistics such as estimated counts are based on sample size and the relative length of the confidence interval. Table 8.2 provides the release rules for the SOLMP, where the domain of interest is either at the national level (e.g., Canada), sub-national domains outside of the territorial capitals, or the territorial capitals.

Table 8.2
Release Rules for SOLMP Estimates

Release Category	Rule		Action
	For proportions[2]	For other statistics[3]	
Rules for estimates at the national level			
A[1] Releasable	n ≥ 180 and CI length ≤ 14 p.p.	n ≥ 180 and Relative CI length ≤ 1.4	Release with no warning; users should use CI as quality indicator
E Releasable with warning	90 ≤ n < 180 or CI length > 14 p.p.	90 ≤ n < 180 or Relative CI length > 1.4	Release with quality warning (letter E); users should use CI as quality indicator
F Not releasable	n < 90 (regardless of CI length)	n < 90 (regardless of CI length)	Suppress the estimate and its CI for quality reasons (letter F)
Rules for estimates at sub-national level			
A[1] Releasable	n ≥ 150 and CI length ≤ 14 p.p.	n ≥ 150 and Relative CI length ≤ 1.4	Release with no warning; users should use CI as quality indicator
E Releasable with warning	75 ≤ n < 150 or CI length > 14 p.p.	75 ≤ n < 150 or CI relative length > 1.4	Release with quality warning (letter E); users should use CI as quality indicator
F Not releasable	n < 75 (regardless of CI length)	n < 75 (regardless of CI length)	Suppress the estimate and its CI for quality reasons (letter F)
Rules for estimates at the territorial capitals level			
A[1] Releasable	n ≥ 60 and CI length ≤ 14 p.p.	n ≥ 60 and Relative CI length ≤ 1.4	Release with no warning; users should use CI as quality indicator
E Releasable with warning	30 ≤ n <60 or CI length > 14 p.p.	30 ≤ n < 60 or CI relative length > 1.4	Release with quality warning (letter E); users should use CI as quality indicator
F Not releasable	n < 30 (regardless of CI length)	n < 30 (regardless of CI length)	Suppress the estimate and its CI for quality reasons (letter F)

1. 'A' is not a quality indicator; it should not be released with the estimate. The 95% confidence interval is the quality indicator.

2. For estimated proportions, n is defined as the unweighted count of the number of respondents in the denominator (not the numerator) of the proportion.

3. For estimated means, n is defined as the unweighted count of the number of respondents that contribute to the estimate including values of zero. For estimated totals and counts, n is defined as the unweighted count of the number of respondents with nonzero values that contribute to the estimate.

Source: 2022 Survey on the Official Language Minority Population.

8.5.2 Release Rules for Differences

To assign a release category for an estimated difference between two estimates, the analyst must first determine the release category of each of the two estimates using the rules described above. Next, the release category of the estimated difference or the estimate of change is assigned the lower release category of the two estimates; this can be specified as follows:

- If one or both estimates are category F estimates, then assign the estimated difference to category F and suppress it.
- Otherwise, if one or both estimates are category E estimates, then assign the estimated difference to category E.
- If both estimates are category A estimates, then assign the estimated difference to category A.

8.5.3 Additional Rules Regarding Confidence intervals

The above release rules should suppress most estimates and confidence intervals of poor quality. There are also two conditions that indicate that a confidence interval is of poor quality. An estimate and its confidence interval should be assigned to release category F if either of the following two conditions are true:

- The lower bound of the 95% confidence interval is equal to the upper bound of the interval; in other words, the confidence interval is of length zero. (Exceptions are if the estimate corresponds to a calibration control total.)
- The lower bound or upper bound of the 95% confidence interval is not a plausible value for the estimate. For example, the lower bound for an estimated proportion is negative.

8.6 Guidelines for Statistical Analysis, Variance Estimation and Constructing Confidence Intervals

To assess the sampling error of estimates, variance estimates are calculated from the Bootstrap weight variables supplied and reported using confidence intervals. The SOLMP uses a complex sample design and estimation method, which means that there is no simple formula for calculating variance estimates. This is why the SOLMP uses a survey-adapted variant of the resampling method called the Bootstrap. One thousand bootstrap replicate weight variables were generated. From these, users get 1,000 Bootstrap estimates by mimicking how the main estimate was obtained from the data and the one survey weight variable. The variance these Bootstrap estimates exhibit serves as an approximation of the variance one would find among the estimates obtained from all possible samples admitted by SOLMP's design.

8.6.1 Statistical packages for statistical analysis and variance estimation

It is necessary to use Bootstrap weight variables to compute correct estimates of the variance for this survey. Several statistical software programs or packages exist to analyze data from complex survey designs, allowing to compute variance estimates using the Bootstrap weight variables. These include for example SUDAAN, WesVar, Stata, R Survey Packages and newer versions of SAS. For more information, please refer to Gagné et al. (2014).

Other standard and/or older statistical analysis software packages - including SPSS, versions of SAS prior to version 9.2 - do not have an integrated procedure to calculate variance estimates from Bootstrap weights when using data based on a complex survey design. These packages should not be used to calculate variance estimates, to construct confidence intervals nor to conduct statistical tests (significance tests, regression analysis, etc.).

SAS version 9.2 and above can calculate variances from Bootstrap weights, as well as other types of replicate weights such as Jackknife and Balanced Repeated Replication (BRR) weights. Various procedures, such as those fitting linear or logistic regression to survey data, accommodate replicate weights. Confidence intervals for medians using replicate weights are only available in SAS version 9.3 and above.

It should be noted that software packages may not explicitly support Bootstrap weights although they do support the BRR method. While the bootstrap and BRR methods differ in the way in which the replicate weights are derived, once obtained they are to be used in the same way by a software to compute variance estimates. This is why users will see SAS code in which varmethod=BRR has been specified even though Bootstrap weights are supplied to the procedure; this workaround is not needed in recent versions of SAS since varmethod=bootstrap is now a valid option.

Important: While certain software packages have procedures that allow for a weight variable to be specified e.g., the WEIGHT statement in PROC FREQ in SAS, it does not mean that specifying the survey weight variable suffices to get valid variances. To illustrate, consider a simple random sample without replacement sample of size 5 with survey weight of 2; from this we understand that the population has 5x2=10 units. However, if the survey sampled data was submitted to PROC FREQ with the survey weight of 2 in its WEIGHT statement, SAS would infer from this the following:

- the sample size is 10, interpreting the survey weight as mere frequencies, that is seeing the sample as consisting of two identical copies of the file with 5 units in it which was specified as input to the procedure;
- the population size is infinite;
- the sample design is akin to simple random sampling with replacement.

These assumptions tacitly made by the software will produce an erroneous statistical assessment of the uncertainty in this survey setting. A proper analysis of the survey data would rather use PROC SURVEYFREQ from SAS survey-suite of procedures, which allows for the Bootstrap weight variables supplied with the data to be used (see SAS documentation regarding REPWEIGHTS statement in conjunction with the option varmethod=Bootstrap).

8.6.2 The Bootstrap Expansion Factor and Fay's Adjustment

As mentioned in Section 7.2 it is critically important to apply an expansion multiplicative factor C of 16 to the standard Bootstrap variance computed from the data to get a proper variance estimate in the context of SOLMP. Most statistical packages allow for such an expansion factor to be used, often under the name of Fay's Adjustment Factor or Fay's Coefficient.

Warning: Because conventions are not uniform across statistical packages, the software documentation must be carefully examined to determine what value to set to its Fay's Coefficient so that the correct value of 16 for the expansion factor C gets used. For example, in SUDAAN Fay's Coefficient needs to be C whereas in SAS it must rather be set to $1 - 1/\sqrt{C}$ to achieve the desired result. Hence, in SAS a Fay Coefficient value of 0.75 is required to get the desired expansion factor of 16.

8.6.3 Confidence intervals

The most commonly used and default choices in most statistical software are the Wald and Student methods.[9] For proportions, the normality assumption is known to be invalid when estimating proportions near zero or one using a sample whose size is not large enough. Three alternative methods of constructing confidence intervals are therefore recommended for proportions: the modified-Wilson interval, the Clopper-Pearson interval and the logit interval (see Korn and Graubard, 1998; Liu and Kott, 2009; Neusy and Mantel, 2016).

Appendix D provides details on how to construct valid confidence intervals using SOLMP's data.

There are options in SAS and SUDAAN to produce confidence intervals using these alternative methods.

The examples below show how Modified-Wilson confidence intervals for estimated proportions can be obtained in SAS and SUDAAN.

1. For older versions of SAS, modified Wilson confidence intervals:

   PROC SURVEYFREQ

   DATA=…. VARMETHOD=BRR (**FAY**=0.75);

   WEIGHT weight;

   REPWEIGHTS bsw1-bsw1000;

   TABLES …. / **CL (TYPE=WILSON ADJUST=NO TRUNCATE=YES)**

2. For recent versions of SAS, modified Wilson confidence intervals:

   PROC SURVEYFREQ

   DATA=…. VARMETHOD=bootstrap;

   WEIGHT weight;

9. Wald and Student's confidence intervals are discussed in Appendix D, as well as in Sections 7.2 and 7.3 of Sampling and Weighting Technical Report, Census of Population, 2021.

REPWEIGHTS bsw1-bsw1000 / **REPCOEFS**=0.016;

TABLES …. / **CL (TYPE=WILSON ADJUST=NO TRUNCATE=YES)**

3. SUDAAN, modified Clopper-Pearson confidence intervals:

PROC CROSSTAB

DATA=…. DESIGN=BRR **SMCONF=50**;

WEIGHT weight;

REPWGT bsw1-bsw1000 / **ADJFAY**=16;

TABLES ...;

Warning: the replication weights MUST be provided to the software because specifying only the variance estimation method (through the statements VARMETHOD for SAS or DESIGN for SUDAAN) will lead to incorrect results. Indeed, in the absence of replication weights most software will undertake implementing the Bootstrap procedure on their own. However, there is NOT a statistical software available that can correctly carry out the Bootstrap method on its own, even in the hands of the savviest Bootstrap user, since a proper derivation of the SOLMP's Bootstrap weights requires survey information that is not provided in the analytical file.

8.6.4 Standardizing the weights

As mentioned already, it is recommended that users use procedures designed for the analysis of data from complex survey designs, which can use weights to produce estimates and can use bootstrap weights to produce variance estimates. Analysis procedures not designed for the complex survey framework may allow for survey weights to be used but not the bootstrap weights. However, these procedures may differ in their definition for the weight, leading to correct estimates but inadequate variance estimates. In the example given in Section 7.0 of a simple random sample of size 100 selected from a population of size 5,000, such a procedure would interpret a sampled unit's survey weight of 50 as a frequency weight: there are 50 such units in the sample. Thus, when presenting the procedure with a data file of 100 units each with a weight of 50, it would consider having a sample of 5,000 in all (that is 50 copies of the survey's sample) taken from a population infinite in size. In addition to ignoring that SOLMP's sample design is complex (notably because the population is finite in size and thus just a fraction larger than the sample), this approach overstates the strength of the data evidence provided by the actual sample.

For analyses such as linear regression, logistic regression and analysis of variance, rescaling or normalizing the weights can make the variance estimates calculated by the procedure more reasonable. The weights for the domain of interest should be rescaled so that the average weight is one (1); this can be accomplished by dividing each weight by the overall average weight for the domain of interest before the analysis is conducted. The rescaling makes the variance estimates more reasonable, but they only consider the unequal probabilities of selection - they do not consider the stratification and clustering of the sample's design. This approach should therefore only be used as a last resort when no procedures that can use bootstrap weights are available; users are warned that the results are approximate.

8.6.5 Use of confidence intervals to determine statistical significance

As discussed in detail in Appendix E, two situations frequently arise in practice:

1. Determining whether the obtained estimate supports or contradicts the hypothesis that the true value is the one postulated, e.g., "Is the true proportion of women exhibiting a certain characteristic 15%?"
2. Determining whether the true values of two groups are different or not, e.g., "Does the true proportion of women exhibiting a certain characteristic differ from that of men?"

A confidence interval can be used to answer such questions. In the first case, it suffices to check whether the hypothesized value lies within the confidence interval associated with the estimate. For example, since 15% is within the confidence interval [10%, 23%], it would be concluded that the data does not reject the hypothesis that the true proportion is 15%.

When comparing estimates between groups, the quantity of interest becomes the difference between the two estimates. Confidence intervals can be used to determine whether this difference is statistically significant. Two methods can be employed to assess statistical significance: comparing the overlap between the confidence intervals of the two estimates and constructing a confidence interval for the difference between the two estimates. They are discussed in the next two sections.

8.6.6 Comparing the overlap between the CIs of the two estimates

When assessing whether the difference in two estimates allows one to conclude that the corresponding parameters of interest are different, one approach contrasts the overlap of their associated CIs. The procedure is as follows. If the two CIs do not overlap, then one can reasonably conclude that the nonzero estimated difference points to a real difference. However, if the two CIs do overlap, no matter how slightly, then one cannot draw any statistical conclusion about the state of the true difference.

This overlap-checking procedure is conservative in the sense that two parameters of interest may be quite different and yet lead to CIs that overlap i.e., the procedure will not correctly detect a difference when one does exist and therefore be lacking power (in statistical parlance). This is because the overlap-checking procedure corresponds to a statistical test based on the largest possible variance for the estimated difference rather than its actual variance. Consequently, for the overlap test to yield a conclusive outcome the gap between the estimates must be larger than the highest potential error. The next section presents a more powerful statistical test, which is based on the actual variance of the difference in the estimates.

8.6.7 Constructing the CI for the difference of the two estimates

The other procedure calls for a confidence interval to be computed directly for the estimate of the difference: if this CI contains zero, then one statistically concludes that no difference exists whereas if it does not contain zero, then one statistically concludes to a nonzero difference. Such a CI can notably be obtained using the bootstrap percentile method described in Appendix D.

8.6.8 Releasing Statistical Information on Gender

As part of the 2022 SOLMP releases, the two-category gender variable will be the standard in data tables and analyses. Users of SOLMP data are encouraged to include the following notes in any tables or analyses they publish when using gender.

Any data tables or analyses released by gender with two categories should include the following notes for Gender, Men+ and Women+:

Gender

Given that the non-binary population is small, data aggregation to a two-category gender variable is necessary most of the time to protect the confidentiality of responses. In these cases, persons in the category “non-binary persons” were randomly distributed in the other two gender categories and are denoted by the “+” symbol.

Men+

This category includes cisgender and transgender men, as well as some non-binary persons.

Women+

This category includes cisgender and transgender women, as well as some non-binary persons.

In analytical products, the use of “men’ or “boys” and “women” or “girls” in the main text instead of “men+” and “women+” is preferred. However, tables, charts and other images included in articles should use “men+” and “women+” and include the applicable notes.

All data tables and analyses involving historical comparisons of the “sex” variable in 2006 and the two-category “gender” variable in 2022 should include the notes above and the following note:

> Gender data were not collected in the 2006 Survey on the Vitality of Official-Language Minorities (SVOLM). Therefore, the sex variable from 2006 and the two-category gender variable from 2022 are combined in this analysis. Although sex and gender refer to two different concepts, the introduction of the gender variable is not expected to have a significant impact on data analysis and historical comparability, given the small size of the transgender and non-binary populations.

This additional note could also be included:

> Although sex and gender refer to two different concepts, gender-related terminology is used in this data table to facilitate reading.

8.6.9 Pooling adults and children samples

The combining of data collected by SOLMP on children and adults requires great caution in order to draw valid and relevant statistical conclusions about this study population. First, the population studied must be precisely described and restricted. For example, it would be incorrect to speak of “children and adults in the official language minority population” since a child’s inclusion in the survey is based on the linguistic profile of his or her parents and not his or her own. In such a case, the target population would be described as “adults in the official language minority population and their children.” In order to gather data that is relevant to this population, it is necessary to only retain children who had at least one adult in the official language minority population in their census family in 2021 (LOSM_C_P = 1) when pooling the adults and children files.

Next, it is necessary to ensure that the data used in the analysis come from survey questions that are identical, or at least sufficiently similar, for both the child and adult segments. In particular, such questions must focus directly on the child, not on the adult responding to the survey on their behalf, otherwise pooling the samples of children and adults will result in an overrepresentation of adults and erroneous estimates for this population.

It is important to note that combining the adult and child files does not result in a sample of related individuals. On the one hand, among the adults in the official language minority population sampled are people who do not have children, or who have children who have not been sampled. On the other hand, although the data for children were collected through adult respondents, these related adults may well not have been sampled, and even if they had been, the data in the files made available would not be able to establish family relationships between the individuals in the adult and child files.

8.7 Differences between the SOLMP and other data sources

Due to the updating of the questionnaire and some differences between the methodology used for the 2022 Survey on the Official Language Minority Population (SOLMP), the 2021 Census and the 2006 Survey on the Vitality of Official-Language Minorities (SVOLM), caution should be exercised when comparing data between these sources. The following sections provide information on elements affecting data comparability, providing data users with important information on factors to consider when performing analyses using 2022 SOLMP and 2006 EVMLO data.

The 2022 SOLMP sample was selected from respondents who provided selected answers to the Census. More detailed information on how Census responses were used to determine the SOLMP target population is provided in section 3 (survey methodology).

8.7.1 Differences between the 2022 SOLMP and the 2021 Census

One of SOLMP's objectives was to contribute to meeting data needs not met by existing data, such as the Census. The Census and the SOLMP are two important sources of information on the official language minority population that complement each other. The SOLMP is based on concepts covered in the Census and includes more in-depth questions to produce more detailed information. For example, Census questions 8 to 10 provide information on language knowledge and use, and questions 12 to 17 on language of instruction at primary and secondary levels. The addition of information from the SOLMP provides an opportunity to learn more about the language skills of adults and children, as well as the language of instruction for post-secondary education for adults and at different levels for children, for example, parents' intentions regarding their children's language of instruction, and their reasons and preferences for choosing their children's language of instruction.

The SOLMP also covers topics or themes that are not covered by the Census. For example, the SOLMP provides information on whether a child in an official language minority situation attended daycare in either of the official languages, and on access to or use of services in the minority official language by adults in the official language minority population. The SOLMP provides an opportunity to obtain more detailed information from children or adults about their experience related to language dynamics, arts, culture and media, sense of belonging, vitality and community participation, and the use of official languages in the public sphere.

Population counts from the 2022 SOLMP for certain sub-populations may differ from those obtained from the Census, even when the Census population universe is restricted to the SOLMP's universe. Weighting ensures that the number of adults in the official language minority population is the same in the Census and the SOLMP, but only for certain combinations of this population, region, language and age groups in the sample strata. Similarly, the number of children will be identical whether calculated from the Census or the SOLMP, but only for certain combinations of region, language and age groups. However, population counts may differ for other sub-populations, which were not controlled for during sampling allocation and post-stratification (see sections 3 and 7).

Furthermore, although some concepts are common to the SOLMP and the 2021 Census of Population, the results may differ slightly between these two sources. For the same person, characteristics reported may, in some cases, differ between the SOLMP and the Census. There are many reasons why responses from these data sources may differ, including the effect of time and proxy reporting. The time elapsed between the 2021 Census and the 2022 SOLMP collection period may explain why some characteristics vary, particularly those that are not fixed over time.

In most regions, data for the 2021 Census was collected using self-enumeration. Questionnaires were completed by respondents, online or via paper and returned by mail. Often, the same person completed the Census questionnaire for all members of the household. This is known as proxy reporting. Data for the SOLMP were collected, in most cases, through self-declaration by the selected adult or parent of the selected child, using an electronic questionnaire. Since the person contacted for the SOLMP is not necessarily the same person who completed the Census questionnaire, responses to similar questions may vary.

8.7.2 Comparability between the 2022 SOLMP and the 2006 SVOLM

The Survey on the Official Language Minority Population (SOLMP) was designed to provide data on current issues and topics that have an impact on official-language minorities, while measuring, to the extent possible, changes in the situation of official-language minority communities since the last similar survey, the 2006 Survey on the Vitality of Official-Language Minorities (SVOLM).

The main new topics addressed in the 2022 SOLMP are the educational attainment of parents of adults and children, parents' intentions regarding their children's language of instruction, adults' hesitation to use the minority official language, perceived discrimination related to linguistic reasons among parents and adults, services to assist immigrants' integration, and the impact of the COVID-19 pandemic on access to and use of services in the minority official language. It is therefore not possible, for these new topics, to make a comparison with the situation that prevailed in 2006.

In addition, the 2006 SVOLM questionnaire has been updated to ensure the relevance of the 2022 SOLMP. Some questions in the 2022 SOLMP remain identical or sufficiently similar to those in the 2006 SVOLM to ensure the

comparability of many results between these two sources (see questionnaires and codebooks for both surveys). However, some of these changes limit the comparability of certain 2022 SOLMP results with those of the 2006 SVOLM. For example, the concept of sex used in the 2006 SVOLM has been replaced by gender in the 2022 SOLMP. In rare cases, significant changes to a module's questions and their structure do not allow comparability over time, as is the case with the 2022 SOLMP health services module. Appendix A provides general indications as to the comparability of the 2006 SVOLM with the main concepts contained in the various modules of the 2022 SOLMP. Some of the content of the 2006 SVOLM was not repeated in the 2022 SOLMP, to limit the response burden on respondents. However, much of the information available in the 2021 Census has been added to the 2022 SOLMP microdata files.

In addition to differences in content, the 2006 SVOLM and 2022 SOLMP samples also contain a few differences that must be taken into account when comparing these two sources. The two main differences, described in the following two paragraphs, are the addition, in 2022, of some children eligible for instruction in the minority official language, in the children's file, and the removal of the additional sample of persons whose mother tongue is neither English nor French, but with French as their first official language spoken, in the adults' file.

The 2022 SOLMP sample includes children eligible for instruction in the minority official language who did not have a parent in the minority official language population. The new questions on language of instruction added to the 2021 Census made it possible to include them, which was not possible in 2006. For this reason, when comparing 2022 SOLMP data on children, these comparisons should only include children who had at least one adult who is part of the official language minority in their Census family in 2021 (LOSM_C_P = 1), so that this universe corresponds to that of children in the 2006 SVOLM.

The main difference between the 2006 SVOLM and 2022 SOLMP adult files is that the 2006 SVOLM file includes a sample of "French first official language spoken (FOLS) allophone" adults who were surveyed in the Montreal area for specific analytical purposes (see section 3.2 of this guide and the SVOLM guide). As this sample was not strictly part of the 2006 target population, these adults must be excluded from the 2006 SVOLM file when making comparisons that include adults from the Montreal region.

Additionally, two language sub-groups of the 2006 SVOLM target population were not included in the 2022 SOLMP sample. These are persons whose mother tongue is neither French nor English, who know both official languages, and who also speak both official languages most often at home (alone or with other languages), and the rare persons whose mother tongues are the majority official language and another language, without the minority official language, and who know only the minority official language. Although these two sub-groups represented less than 1% of the target population of the 2006 SVOLM (see tables B.1 and B.3 of the 2006 SVOLM methodological guide), it is nevertheless preferable to remove them when making comparisons with the 2022 SOLMP.

Other differences relate to selected geographical regions. As indicated in the summary table in Appendix A, it is not possible to compare the estimates for the territories from the 2006 SVOLM with those from the 2022 SOLMP, since in 2022 only the territorial capitals were included in the sample (see section 3.1 of this guide). Other differences between the geographic regions of the 2006 SVOLM and the 2022 SOLMP relate to the regions of New Brunswick, which are slightly different in the 2022 SOLMP. The North region of New Brunswick now includes the two 2021 census subdivisions (CSDs) with a strong francophone presence, Alnwick and Neguac. The Southeast region of New Brunswick now also includes the 2021 census subdivisions (CSDs) of Rogersville (parish and village) and Hardwicke, which includes the French-speaking community of Baie-Sainte-Anne. Comparisons with this province's regions in the 2006 SVOLM should be made with caution, particularly for the "rest of New Brunswick" region.

Finally, other differences between the 2006 SVOLM and 2022 SOLMP samples are minor, and generally have a negligible effect on overall comparisons between these two sources. For example, both postcensal surveys excluded permanent residents from their samples; however, the method of identifying permanent residents has changed since 2006.

In addition, it is important to note that the 2006 SVOLM child file does not have a unique identifier for each child, so two variables (sampleid and childid) must be used to sort the files before linking the bootstrap weights.

Appendix A – 2022 SOLMP content themes and concepts and comparability with the 2006 SVOLM

Table A
2022 SOLMP content themes and concepts and comparability with the 2006 SVOLM

Themes	2022 Questionnaire Modules	Masterfiles	Main Content and Concepts	Comparability with the 2006 SVOLM
Sociodemo-graphic	Respondent information (NAM)	Adults & Children	Age of adult or child's parent as of May 16, 2022.	Comparable age groups: 18-24; 25-44; 45-64; 65+.
Sociodemo-graphic	Respondent information (NAM2)	Children	Age of child as of May 16, 2022.	Comparable age groups: 1-4; 5-11; 12-17.
Sociodemo-graphic	Respondent information (ANU)	Adults & Children	Province of residence as of May 16, 2022.	Generally comparable with 2006 except for territories.
Sociodemo-graphic	Gender of targeted respondent (GDR)	Adults & Children	Gender of adult or child (see GND variables in codebooks)	New in 2022. Sex was used in 2006.
Sociodemo-graphic	Respondent information (LAN)	Adults	Main Language characteristics.	Comparable with 2006.
Sociodemo-graphic	Household composition (HHC)	Adults & Children	Household composition: number of persons, their age, gender, relation to respondent.	Expanded to all members of the household in 2022.
Sociodemo-graphic	Household composition (MS)	Adults & Children	Marital status of respondent, or of child's parent, and their partner.	Comparable with 2006.
Sociodemo-graphic	Partner of Respondent (SP)	Adults & Children	Partner's relation to the child, language characteristics and birthplace.	Generally comparable, one question removed in 2022.
Sociodemo-graphic	Parent block (PAR)	Adults & Children	Adult's or parent's father or mother, birthplace, language, and education.	Generally comparable with 2006 except for education.
Sociodemo-graphic	Children in the Household (CHD)	Adults & Children	Main language characteristics of each child (< 18) living in the household.	Expanded to all children in the household in 2022.
Language	Language skills (LSK)	Adults & Children	Adult's or child's language skills. Linguistic discrimination is new in 2022.	Updated module. Some comparability with 2006.
Language	Language skills of parent of child (LSP)	Children	Language skills of the child's parent. Linguistic discrimination is new in 2022.	Updated module. Some comparability with 2006.
Education	Education of respondent (EDU)	Adults	Postsecondary institution attended, location, language and reasons for language.	Generally comparable with 2006.
Education	Education of respondent (EDU)	Children	Child's education: institution, language, parent's satisfaction, reasons and preferences.	Updated module. Some comparability with 2006.
Education	Educational intentions (INT)	Children	Parent's educational Intentions for their child and reasons for such intentions.	New module in 2022. No comparability with 2006.
Education	Early childhood services (CHS)	Children	Child's daycare language(s), parent's reasons and preferences.	Updated module. Some comparability with 2006.
Language	Linguistic trajectory from childhood to adulthood (TRA)	Adults & Children	Linguistic trajectories from childhood to adulthood of the adult or the parent.	Updated module. Some comparability with 2006.
Language	Linguistic dynamic in child's family (DYN)	Children	Child's language spoken at home, with family members, with friends, and at school.	Updated module. Some comparability with 2006.
Identity	Sense of belonging and subjective vitality (VIT)	Adults & Children	The adult's or parent's linguistic identity and perception of the linguistic situation.	Updated module. Some comparability with 2006.
Participation	Civic participation (CIV)	Adults	Adult's participation in volunteer work, community organizations, and associations.	Updated module. Some comparability with 2006.
Language	Language use in the public sphere (PUB)	Adults	Adult's, child's and parent's use of languages in the public space.	Mostly new content. Limited comparability with 2006.
Services	Government (GOV)	Adults	Adult's preference, access, use and satisfaction with language of government services.	Updated module. Some comparability with 2006.

Table A
2022 SOLMP content themes and concepts and comparability with the 2006 SVOLM

Themes	2022 Questionnaire Modules	Masterfiles	Main Content and Concepts	Comparability with the 2006 SVOLM
Services	Justice (JUS)	Adults	Adult's access, use and satisfaction with language of police and justice services.	Updated module. Some comparability with 2006.
Services	Health (HLT)	Adults	Adult's importance, access, use and satisfaction with language of health care or services.	Updated module. No comparability with 2006.
Services	Immigrant services (IMS)	Adults	Language of services obtained at arrival in Canada by adult immigrants.	New module in 2022. No comparability with 2006.
Culture	Art, culture and media (ACM)	Adults & Children	Adult's or child's language of participation in arts, culture, media activities.	Updated module. Some comparability with 2006.
Mobility	Geographic mobility (GEO)	Adults & Children	Adult's or child's birthplace, at 15 years, and at collection time. Reasons for migrations.	Updated module. Some comparability with 2006.
Work	Labour market (LAB)	Adults & Children	Adult's or parent's labour force activity, the type and location of their job.	Updated module. Some comparability with 2006.
COVID	COVID-19 (COV)	Adults & Children	Adult's or parent's change of access to services in official languages due to COVID-19.	New module in 2022. No comparability with 2006.

Reference: Survey on the Vitality of Official-Language Minorities (SVOLM), 2006

Source: 2022 Survey on the Official Language Minority Population

Appendix B – Examples of estimate and confidence interval calculations

This appendix provides an example of how to calculate a point estimate and its confidence interval (CI) from the Survey on the Official Language Minority Population (SOLMP). Note that to determine the confidence interval (CI) of the estimates, statistical software or packages that can calculate variance estimates from bootstrap weights must be used. In the following examples, SAS is used.

Estimating the percentage of adults from the official language minority population who had experienced a situation in which they hesitated to use the minority official language in the five years preceding the survey (see section 2.2 Linguistic insecurity in the survey analytical report Situation of English-speaking populations in Quebec and French-speaking populations in Canada outside Quebec: Results of the 2022 Survey on the Official Language Minority Population**).**

Start by consulting the language skills (LSK) module section in the topical Index at the end of the codebook to find the corresponding variable (LSK_55). The codebook entry for this variable provides its abbreviated concept, question wording, universe and unweighted frequencies. This confirms that variable LSK_55 indicates, among all respondents, those who hesitated to use the minority language in the past five years.

The first step in producing the population estimate and its confidence interval (CI) is to link the analytical file with the bootstrap weights file (see section 7 for more on the bootstrap weights).

```
data SOLMP_ADULTS;

    merge SOLMP.solmp_adt_m SOLMP.solmp_adt_bsw;

    by masterid;

run;
```

To calculate the required percentage and its confidence interval (CI), a frequency table should be produced using the SURVEYFREQ procedure, as shown in the following SAS code example.

```
PROC SURVEYFREQ data=SOLMP_ADULTS varmethod=bootstrap;

      WEIGHT weight;

      REPWEIGHTS bsw1-bsw1000 /repcoefs=0.016;

      TABLES LSK_55 / clwt cl (type=wilson adjust=no truncate=yes);

      WHERE LSK_55 IN (1,2);

RUN;
```

The code above specifies the bootstrap option as method to calculate variance estimates, using the variable ‘weight’ as weight and the bootstrap weights bsw1 to bsw1000. For the SOLMP, it is extremely important to use the appropriate multiplicative factor (repcoefs=0.016), often called the “Fay adjustment factor,” for any sampling error estimate such as the CI (see section 8.6.2).

The TABLES line specifies the variable of interest and the “cl” option to obtain the confidence interval (CI) of the estimates only for the respondents who answered yes or no to the question, as specified by the WHERE line.

The rounded results are presented in the table below.

Table B1
Estimates and confidence intervals for adults, Canada

	Estimated population size	95% confidence intervals for population size		Estimated proportion	95% confidence intervals for proportion	
Hesitation to use the minority official language	rounded weighted number	from	to	percent, based on unrounded weighted numbers	from	to
Yes	563,120	539,190	587,040	27.3	26.2	28.5
No	1,499,100	1,475,160	1,523,039	72.7	71.5	73.8
Total	**2,062,220**	**2,059,680**	**2,064,750**	**100.0**	**...**	**...**

... not applicable

Note: The SAS SURVEYFREQ procedure calculates the Wald confidence limits for weighted counts and the modified Wilson confidence limits for percentages. See SAS documentation for more details.

Source: 2022 Survey on the Official Language Minority Population.

According to this table as stated in section 2.2 of the SOLMP report, "In Canada, approximately one-quarter (27%) of adults in the OLM population had experienced […] over the previous five years […] a situation in which they hesitated to use the minority official language." (Pépin-Filion et al., 2024).

It is important to note that to obtain the percentage in the table, according to the rounding guidelines, unrounded weighted figures from SAS must be used in the numerator and denominator (i.e., 563,116/ 2,062,216 = 27.3%). The 95% CI around this estimate ranges from a lower bound of 26.2% (rounded) to an upper bound of 28.5% (rounded).

Please note that the unweighted figure in the numerator (obtained using PROC SURVEYFREQ and not shown here) on which this estimate is based is well above the minimum value of 10 required for a statistic to be released under confidentiality rules. Please refer to section 8.1 of this document for more information.

It is also necessary to ensure that the estimate can be disseminated, with or without a warning, in accordance with the dissemination quality rules. Since the domain of interest is at the national level (all of Canada), the rules in Table 8.2 must be used specifically. Here, the unweighted number of units in the denominator (not shown here) is ≥ 180 and the confidence interval (CI) length is 2.3 percentage points (p.p.), which is less than 14 p.p. Thus, the 27.3% proportion and its CI (26.2 % to 28.5%) can be released without a warning (release category A), and users should use the CI as a quality indicator.

Determining whether the observed difference between two estimates is statistically significant

Once the 95% CIs have been established, it is relatively simple to determine whether the difference between two estimates is statistically significant. If the two intervals overlap, then it cannot be concluded that the underlying population quantities (for instance, specific proportions in the population for two groups of people) being estimated are different. If the two intervals do not overlap, however, it can be concluded that the underlying population quantities being estimated are different. In more technical terms, the null hypothesis that there is no difference between the underlying population quantities being estimated, at the 5% level, can be rejected. As mentioned previously in section 8.6.5, this is a conservative approach and not an exact one: the actual rejection rate is lower than the 5% targeted here. One way of performing an exact test would be to create a CI for the difference between the two proportions, or to use the SAS DIFFMEANS procedure.

Continuing with the previous example, suppose a user wants to determine whether there is a statistically significant difference between the percentage of English-speaking adults who hesitated to use English in Quebec than French-speaking adults who hesitated to use French in Canada outside Quebec. The following table was created the same way as above, adding a second variable "IN_QUE" and the option "row" to the TABLES line to produce a crosstabulation comparing the hesitation to use the minority language in Quebec and outside Quebec:

```
PROC SURVEYFREQ data=SOLMP_ADULTS varmethod=bootstrap;

    WEIGHT weight;

    REPWEIGHTS bsw1-bsw1000 /repcoefs=0.016;

    TABLES IN_QUE * LSK_55 / clwt cl row (type=wilson adjust=no truncate=yes);

    WHERE LSK_55 IN (1,2);

RUN;
```

Table B2
Estimates and confidence intervals for adults, Quebec, Canada outside Quebec

	Estimated population size	95% confidence intervals for population size		Estimated proportion	95% confidence intervals for proportion	
Hesitation to use the minority official language	rounded weighted number	from	to	percent, based on unrounded weighted numbers	from	to
Quebec						
Yes	324,280	304,880	343,690	30.5	28.7	32.3
No	720,350	720,350	759,310	69.5	67.7	71.3
Total	**1,064,110**	**1,061,690**	**1,066,530**	**100**	**...**	**...**
Canada outside Quebec						
Yes	238,840	225,760	251,920	23.9	22.6	25.3
No	759,270	746,200	772,340	76.0	74.7	77.4
Total	**998,110**	**997,300**	**998,910**	**100**	**...**	**...**
Total						
Yes	563,120	539,190	587,040	27.3	26.2	28.5
No	1,499,100	1,475,160	1,523,039	72.7	71.5	73.8
Total	**2,062,220**	**2,059,680**	**2,064,750**	**100**	**...**	**...**

... not applicable

Note: The SAS SURVEYFREQ procedure calculates the Wald confidence limits for weighted counts and the modified Wilson confidence limits for percentages. See SAS documentation for more details.

Source: 2022 Survey on the Official Language Minority Population.

According to this table, 30.5% of English-speaking adults hesitated to use English in Quebec and the corresponding 95% confidence interval is 28.7% to 32.3% (figures rounded to one decimal place) compared to 23.9% of French-speaking adults hesitated to use French in Canada outside Quebec with the CI of 22.6% to 25.3%.

To assess whether the observed difference between the two estimates is statistically significant, the two 95% CIs must be compared:

Canada outside Quebec:	Quebec:
22.6% to 25.3%	28.7% to 32.3%

Since the two intervals do not overlap, it can be said, at the 5% significance level, that the proportion of English-speaking adults who hesitated to use English in Quebec was significantly different from the proportion of French-speaking adults who hesitated to use French in Canada outside Quebec. See Section 8.6.5 for further explanations on the use of CIs to determine statistical significance.

Therefore, it's possible to state, as in the SOLMP report, that the: "proportion was slightly higher in Quebec (31%) than in Canada outside Quebec (24%)." (Pépin-Filion et al., 2024).

Appendix C – A non-technical primer to survey methodology

This appendix offers a broad and non-technical introduction to survey methodology allowing to better understand the specific aspects of the SOLMP described in this document. A more comprehensive and yet still accessible account of the survey process can be found in Statistics Canada Survey Methods and Practices.

For the SOLMP a list of all persons of interest, called the (sampling) frame, was created using the data collected by the 2021 Census of Population, which means it is a post-censal survey. The frame was split into mutually exclusive sub-groups called strata, the sampling building blocks from which the analytical domains of interest are made. In each stratum a random sample was drawn; pieced together, these samples form SOLMP's (main) sample, which identifies the persons to be contacted during data collection activities.

Once data from sampled persons have been collected, verified and anonymized, they enter the analytical file alongside the survey weight variable which encapsulates the relevant survey design information needed to make statistical statements (called inferences) about the whole population of interest. In essence, survey inference is all about extrapolating from the part to the whole, since there rarely is any analytical interest in the sampled persons per se as a subset of the population.

Typically, the object of an inference is a certain quantity of interest in the population (called a parameter of interest). The proportion of persons in the population that share a given characteristic is an example of a parameter. The value of the parameter of interest would be known if only exact data were collected from all persons of interest in the population. However, in practice there will be data only for a subset of these persons, and so the parameter's value is unknown. The question then becomes: how can the collected data be used to produce a plausible value (called an estimate) for the parameter of interest? And once an estimate has been obtained, how plausible or accurate is it?

To extract an estimate from the mass of information contained in the analytical file, a mathematical device called an estimator is used. There are two components to an estimator: the relevant data to be summarized into one value – the estimate – and the survey weight variable, which will ensure that the sample-based estimate aligns with the population's parameter of interest.

Survey error is the numerical difference between the one estimate obtained from the data and the corresponding true value of the parameter of interest. While survey-error-free situations exist in textbooks, in practice there is always survey error to contend with (survey errors exists in a census – a survey where the sample matches the whole frame – because the responses provided may not be exact, and the frame may not match the population of inferential interest and thus present coverage issues).

While there is no way of knowing for sure in practice how big the survey error is – that is, just how far the one estimate obtained from the collected data stands to the parameter of interest – there is a way of knowing how likely the estimate is to be close to the true value, and this involves assessing the estimator's bias and variance.

Consider this fictional numerical example which sets up the table for a description of the bias and variance. Suppose that the true proportion of interest is 15% in a population and the estimate based on the collected data from a randomly selected sample is 12.6%, which is seen to stand 2.4 percentage points (pp) below the true value. Had another sample been selected, perhaps the estimate would have been 13.4%, which itself stands 1.6 pp below the true value. Observe also that these two estimates, 12.6% and 13.4%, are 0.8 pp apart from one another.

With just two samples, the following survey error patterns are emerging: 1) the estimates obtained here tend to fall below the true value, and 2) they do not match but rather stand some distance away one from the other. The first pattern hints to bias in the estimates – which is their tendency to land on the same side of the true value – whereas the second speaks to their variance, which is their tendency to disagree among themselves as to what the plausible value put forward ought to be considering the data. To formally establish the bias and the variance of an estimator, we would need to examine all possible samples and their estimates, not just two. Since in practice we only have data from one sample, assessing bias and variance promises to be challenging, but it can be done by resorting to bias-mitigating best practices and using a survey-suited variance estimation method.

Naturally, we want survey errors overall to be as small as operationally feasible. But since there are two components to contend with – bias and variance – reducing survey errors becomes a bias-variance trade off. Various other accounts of bias and variance in a survey context exist, including the following[10].

Best survey-conducting and inference-making practices aim at minimizing the bias, leaving then only the variance to be assessed. There are many reasons for pursuing this specific bias-variance trade off. First, while bias is a statistical term, it is also a negative-sounding word in everyday language. Second, bias is unaffected by an increase in sample size: chances are, using more data from the same source will only strengthen the hold the bias already has on the estimates. In contrast, the variance does diminish when the sample size is increased, which presents the survey team with a convenient way of reducing survey error. Finally, statistical bias is notoriously difficult to assess in practice since the unknown true value of interest is needed to establish the direction and magnitude of the survey error. Again, in contrast, the variance usually can be assessed since it solely depends on how the estimates behave as a whole – how close or far apart they stand from one another – and not on where they collectively stand with respect to the (unknown) parameter of interest.

In some cases, survey error can largely be prevented by adopting best survey practices. For instance, before data collection started, SOLMP's questionnaire went through an extensive and rigorous review to ensure that questions were clearly posed and consistent in both French and English. However, in other cases such as sampling and nonresponse, survey error will occur despite the survey team's best efforts and practices. The goal then is to mitigate survey error as much as possible, which the survey team does by relying on proven and well-tested methodologies when producing the analytical file. Then, by using the survey weight variable provided in the analytical dataset, analysts work to eliminate bias from their inferences, thereby converting whatever remains of the survey error into variance which is to be assessed using the Bootstrap weight variables also supplied.

To illustrate the bias-variance trade-off that comes from using the survey weight variable, consider the following fictional numerical example. Suppose that 235 persons in a population of one thousand share a given characteristic. To estimate this count of interest, a Simple Random Sample Without Replacement sample of 500 persons is drawn, which is half the size of the population. Consider the unweighted estimator which simply returns as an estimate of the number of persons of interest found in the sample, which is say 115 for this one sample. Even without knowing the true count the bias of the unweighted estimator is evident, since it only counts the persons of interest in half of the population: the other half is completely unaccounted for. Now consider the weighted estimator that effectively doubles the sample count; this factor of 2 corresponds to the survey weight in this case. While not error free, the weighted estimate of 230 is at least comparable in magnitude to the true value of 235.

Hence, the survey weight variable has reduced the bias the unweighted estimator first presented (actually, one can show that the weighted estimator is unbiased in this case). But because a weighted count is double its unweighted counterpart, inevitably the weighted estimates will stand further apart from one another than the unweighted estimates themselves did. Hence, getting rid of the bias brought on additional variance. The use of the survey weight variable constitutes a sound bias-variance trade-off because all the remaining survey error there is in a weighted count is now in variance form, which can be assessed using the Bootstrap weight variables supplied.

Why is there just one survey weight variable in the analytical file, but a thousand Bootstrap weight variables provided with the data? Ideally, the exact variance would be assessed by repeatedly drawing samples from the population, but this is not doable since a survey can only afford to collect data for one such sample, and the survey weight variable is attached to that one sample. The Bootstrap procedure carried out by the survey team involves cleverly selecting 1,000 samples from SOLMP main sample in such a way that the variability observed in the ensuing 1,000 Bootstrap estimates will approximate the variance that would be observed among all estimates.

Once the variance has been assessed using the Bootstrap weight variables supplied, it is to be reported through a confidence interval (CI) rather than a coefficient of variation (CV). Indeed, experience and investigative work (e.g., Neusy and Mantel, 2016) have shown that CVs can be misleading when reporting on the accuracy of estimates of proportions or counts. Section 8.6 and Appendix D provides more information about the computation and interpretation of confidence intervals as accuracy measures.

10. Variance and bias and Statistics 101: Statistical Bias

Appendix D – Methods used to construct confidence intervals for the 2022 postcensal surveys

This appendix provides general information about confidence intervals as well as details about the methods used to construct the confidence intervals disseminated by Statistics Canada for the 2022 postcensal surveys, namely the Canadian Survey on Disability (CSD), the Survey on the Official Language Minority Population (SOLMP), the Indigenous Peoples Survey (IPS) and the Nunavut Inuit Supplement (IPS–NIS). These methods are the modified Wilson confidence interval for proportion-type statistics and the bootstrap percentile method for other statistics. This document also provides high-level information about the underlying assumptions and the properties of these methods. Users interested in theoretical details are invited to consult the referenced papers.

About confidence intervals

Confidence intervals are provided along postcensal surveys estimates so that users can perform valid statistical inferences. As discussed further in Appendix E, statistical inference is the process of drawing conclusions about the population based on data collected from survey respondents. Survey estimates produced from a sample are subject to uncertainty because of the sampling process—different estimates could have been obtained if a different sample had been selected—and of non-response. This uncertainty must be considered to do proper statistical inference.

As described in the documentation of each survey, variance estimates are produced to quantify the uncertainty of estimates. These variance estimates can be used to produce variance-based quality measures such as standard errors, coefficients of variation and confidence intervals. Among these measures, confidence intervals were chosen to accompany estimates for the 2022 postcensal surveys because they convey that there is uncertainty around the estimate by providing a range of possible values. They also provide an objective measure of the sampling error in a form that is easy to interpret, since wider confidence intervals are naturally associated with greater uncertainty. Moreover, confidence intervals are appropriate for all types of estimates, which is not the case for some other variance-based quality measures.

A confidence interval is associated with a confidence level. The confidence level describes the degree to which one can be confident that the true population parameter is contained in the confidence interval. The confidence level is often written as $1-\alpha$, where α is the significance level of the corresponding hypothesis test. For the postcensal surveys, the confidence level is set to 95% and α is thus equal to 0.05. The confidence level corresponds to the expected coverage of confidence intervals. In other words, if the process that generated the sample data was repeated a very large number of times and corresponding confidence intervals were constructed for each estimate using the same method, the proportion of these confidence intervals containing the true population value should be close to the stated (or nominal) confidence level. Using simulation studies, confidence intervals methods can be evaluated to ensure that their actual coverage is close to their nominal confidence level.

Several different methods can be used to construct confidence intervals. The most commonly used and default in most statistical software are the Wald and the Student's methods.[11] Both of these methods assume a known distribution for the estimator, i.e., the statistical function used to produce estimates. The distribution of the estimator can be seen as the geometrical shape that would appear if all possible samples were drawn from the population and estimation was carried out for each of them, yielding a different estimate each time. While the Wald and the Student's confidence intervals are easy to compute and generally valid, they rely on assumptions that do not hold in some contexts, which can lead to undercoverage.

Other methods, such as modified Wilson for proportions, make different assumptions of the distribution of the estimator. Their improved coverage makes them a better choice for valid statistical inference. Interestingly, the bootstrap percentile confidence interval is obtained in a completely different way. Rather than assuming a shape for the distribution of the estimator, the bootstrap method uses the empirical distribution that arises from the estimates computed from each of the bootstrap replicates. Both methods are described in more detail in the next sections.

11. Wald and Student's confidence intervals are discussed in Sections 7.2 and 7.3 of Sampling and Weighting Technical Report, Census of Population, 2021.

Modified Wilson confidence interval

For the postcensal surveys, this method is used for all proportion-type statistics (e.g., proportions,[12] percent distribution,[13] shares[14]).

The modified Wilson confidence interval method was chosen because of its generally superior coverage for proportion-type estimators and for its practicality of implementation. The method is based on the Wilson confidence interval for a simple random sampling with replacement (SRSWR) sample design (Wilson 1927). For the postcensal surveys, a modified version of the Wilson confidence interval that has been adapted to complex sample designs is used (Kott and Carr 1997). Theoretical developments and extensive simulation studies have shown that this method has good properties in most situations. It performs better than the Wald and Student's confidence intervals in situations where those confidence intervals exhibit undercoverage for proportion-type statistics because the assumptions they rely on do not hold (Neusy and Mantel 2016; Statistics Canada 2023).

Method

The lower bound and the upper bound of a 95% modified Wilson confidence interval for a proportion-type statistic p are given by:

$$\text{lower bound} = \frac{\hat{p} + z^2/2n_e}{1 + z^2/n_e} - \frac{z\sqrt{\hat{p}(1-\hat{p}) + z^2/4n_e}}{\sqrt{n_e}\left(1 + z^2/n_e\right)}$$

$$\text{upper bound} = \frac{\hat{p} + z^2/2n_e}{1 + z^2/n_e} + \frac{z\sqrt{\hat{p}(1-\hat{p}) + z^2/4n_e}}{\sqrt{n_e}\left(1 + z^2/n_e\right)}$$

where:

- $\hat{p}$ is the estimate of p;
- z is the $1-\alpha/2 = 97.5$th percentile of the standard normal distribution;
- $n_e = \min\left(\frac{n}{deff(\hat{p})}, n\right)$ is the effective sample size;
- $deff(\hat{p}) = \frac{\hat{V}(\hat{p})}{\hat{p}(1-\hat{p})/n}$ is the estimated design effect of $\hat{p}$ with respect to the SRSWR sample design;
- n is the sample size (i.e., the unweighted count of respondents) in the denominator of the proportion;
- $\hat{V}(\hat{p})$ is the estimated variance of $\hat{p}$.

When $\hat{p} = 0$ or $\hat{p} = 1$, the value for n_e is undefined. In this case, the bounds for the Wilson confidence interval are computed as:

- For $\hat{p} = 0$, the confidence interval bounds are $\left(0, \frac{1}{1 + n/z^2}\right)$
- For $\hat{p} = 1$, the confidence interval bounds are $\left(\frac{1}{1 + n/z^2}, 1\right)$.

12. A proportion is a relationship between two quantities, often expressed as a fraction. It usually represents the comparison of one part to the whole.
13. Percent distribution refers to the breakdown of a total into its constituent parts, expressed as a percentage of the whole.
14. A share typically refers to the portion or percentage of a particular category or segment within a dataset. In the context of income distribution, a share refers to the portion or percentage of the total income held by a specific group.

Properties

In addition to achieving better coverage than the Wald and Student's intervals for small sample sizes or when the population parameter is near zero or one, the modified Wilson confidence interval for a proportion has the desirable range-preserving property, meaning that its lower bound is never less than zero and its upper bound is never greater than one. Since proportions cannot take on values outside of the interval between zero and one, it is reasonable that confidence intervals for proportions would exclude negative values and values greater than one.

It should also be noted that, unlike the Wald and Student's intervals, the modified Wilson confidence interval for proportions may be asymmetric, meaning that the estimate will not be exactly at the centre of the interval. The asymmetry is small when the effective sample size is large or when the estimated proportion is near 0.5.

Much like the Wald and Student's confidence intervals, the modified Wilson confidence interval for proportions may suffer from some undercoverage, particularly when the sample size is very small, the value of the proportion is near zero or one, or there is high correlation between members of the same household. However, the modified Wilson method generally achieves nominal coverage rates in extreme situations, compared with the Wald and Student's methods. It generally maintains coverage as good as or better than the Wald and Student's methods in those situations.

Bootstrap percentile confidence interval

For the postcensal surveys, this method is used for all statistics except the proportion-type statistics.

The bootstrap percentile method was chosen because of its straightforward way to derive confidence intervals for various and complex estimators of population parameters. Estimation is repeated with each bootstrap replicate weight and the resulting estimates are used to approximate the distribution of the estimator instead of making assumptions on its shape. The appropriate percentiles of this approximate distribution are then used to delimit an area around the estimate that corresponds to a 95% confidence interval. Theoretical developments and simulation studies have shown that this method has good performance in most situations and can be applied for complex estimators (Efron and Tibshirani 1986; Tibshirani 1984).

Method

The lower bound and the upper bound of a 95% bootstrap percentile confidence interval for a non-proportion-type statistic Y are given by:

$$\text{lower bound} = \hat{Y} + \sqrt{R}\left(\hat{Y}_{(LB)} - \hat{Y}\right)$$

$$\text{upper bound} = \hat{Y} + \sqrt{R}\left(\hat{Y}_{(UB)} - \hat{Y}\right)$$

where:

- $\hat{Y}$ is the estimate[15]
- R is the bootstrap adjustment factor[16]
- $\hat{Y}_{(LB)}$ is the LB^{th} non-missing ascending bootstrap estimate
- $\hat{Y}_{(UB)}$ is the UB^{th} non-missing ascending bootstrap estimate
- $LB = \frac{\alpha}{2} \times B$
- $UB = \left(1 - \frac{\alpha}{2}\right) \times B$

15. $\hat{Y}$ can be a total, a quantile, the difference of two estimates, etc.
16. *R* is often called the Fay adjustment factor for postcensal surveys. For CSD, IPS and SOLMP, R = 16.

- α is the significance level
- B is the number of replicates (excluding the number of replicates where $\hat{Y}_j = .$, where $j = 1, ..., B$)[17].

If LB and UB are not integers, then $\hat{Y}_{(LB)}$ and $\hat{Y}_{(UB)}$ is the average of the two contiguous ascending bootstrap estimates. For the postcensal surveys, the values LB and UB are respectively equal to 25 and 975.

For postcensal surveys, it is critical to apply the bootstrap adjustment factor R, otherwise the margin of error associated with the estimate will be underestimated. In other words, the length of the confidence interval will be underestimated if the Fay adjustment factor is omitted in the calculation of the lower and upper bounds.

Properties

Unlike the Wald and Student's intervals, the bootstrap percentile confidence interval is transformation-respecting, which means that the method is valid if there exists a monotone transformation of the estimator that normalizes the distribution of the estimator. This transformation does not have to be known; it only has to exist. In the case of the Wald and Student's intervals, such a transformation would have to be explicitly specified.

Moreover, like the modified Wilson confidence interval, the bootstrap percentile interval is range-preserving, which means that the method always produces intervals that fall within the allowable range of values for the parameter.

Much like other confidence intervals, the bootstrap percentile confidence interval does not perform well when the estimator is biased. Bias in the bootstrap distribution leads to bias in the confidence interval. It is also known that the coverage of the bootstrap percentile confidence interval tends to be lower than the nominal rate when the sample size is small. Nevertheless, it usually achieves better balance in the left and right sides compared with the Wald and Student's intervals (Efron and Tibshirani 1994).

17. There are B=1,000 bootstrap replicates for each postcensal survey.

Appendix E – A Primer of Complex Survey Data Analysis

E1. Introduction

Analyzing complex[18] survey data can appear a daunting task, with many pitfalls awaiting the unsuspecting analyst. Indeed, the classical statistical techniques and software users may be familiar with are generally not suited for use with complex survey data and must be adapted. For example, despite its name, the survey weight has little to do with the weight variable featured in a weighted linear regression, a technique intended for non-survey data that is commonly found in statistical packages.

Analysts must use a software with survey-adapted procedures to carry out their analyses on the SOLMP's data using the survey weight and Bootstrap weight variables supplied. In the case of SAS, this means resorting to its SURVEY-suite of procedures rather than to its classical procedures e.g., use PROC SURVEYREG instead of PROC REG to fit a linear regression model to the SOLMP's data.

Fortunately, many options exist for an analyst to get better acquainted with survey data analysis, including books such as Lohr (2021), Heeringa et al. (2020), Lumley (2010) and Lewis (2016), as well as online material and tutorials such as these[19]. Also, analysts are strongly encouraged to seek guidance and advice from an experienced analyst or a survey statistician throughout the analytical process.

It is an analytical process indeed, as there is more to analyzing survey data than determining which methods and software to use; for instance, Heeringa et al. (2020) identify and discuss the following six steps that an analytical plan ought to cover to yield a statistically valid analysis of survey data that is meaningful to both the social and research communities:

1. Defining the problem and stating the objectives
2. Understanding the sample design
3. Understanding design variables, underlying constructs and missing data
4. Analyzing the data
5. Interpreting and evaluating the results of the analysis
6. Reporting of estimates and inferences from the survey data.

Whether an analytical plan is carried out solely for exploratory purposes or to support data-driven decision making, the survey weight and Bootstrap weight variables must be used with the SOLMP's data to deal with survey error.

This Appendix provides a brief introduction to statistical testing, beginning with a discussion of the role to be played by the survey weight and Bootstrap weight variables in any release of statistical information from the SOLMP's data.

E2. Managing survey error through the survey and Bootstrap weight variables

Survey error refers to the difference between an unknown quantity of interest e.g., the proportion of persons in the population with a certain characteristic, and its estimate based on survey data. Because data are collected only for a subset of all persons of interest (e.g., the respondents from a sample of the population) and that a respondent's survey experience may not yield the exact answers to the questions asked, in practice any estimate is derived from incomplete and imperfect information. Hence, despite the best efforts and practices deployed by the expert survey team, there is always survey error to contend with.

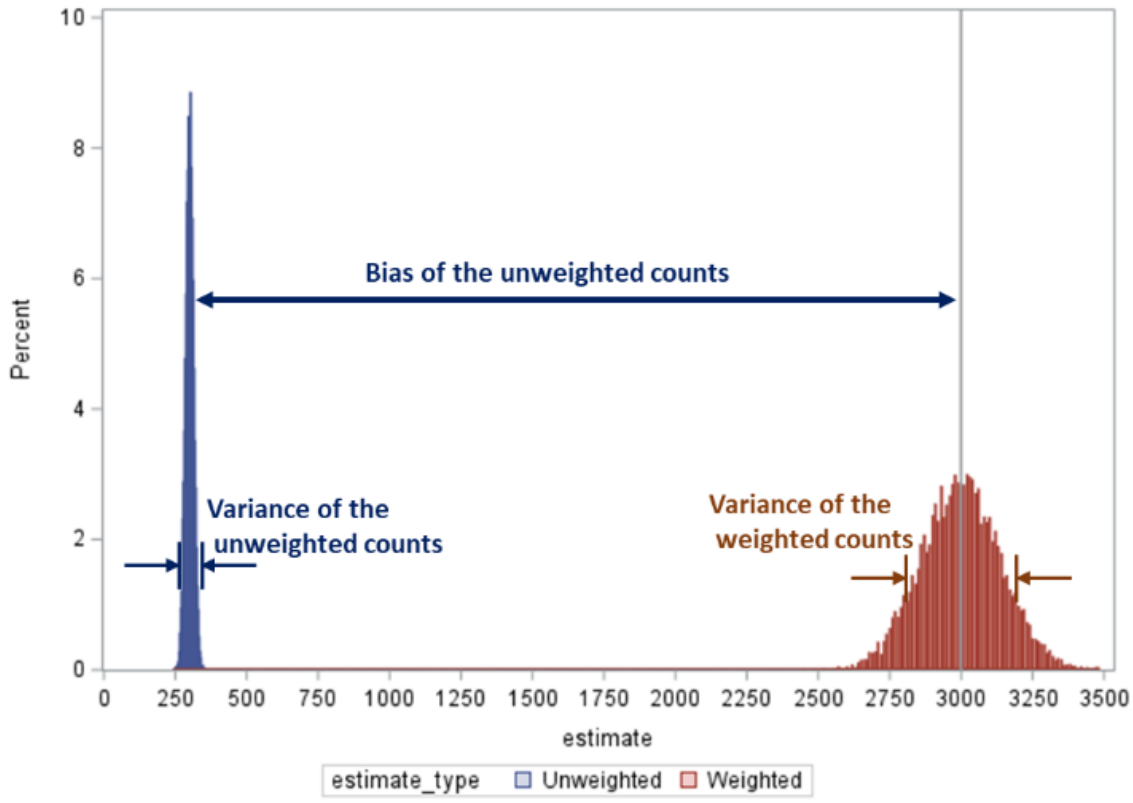
Survey error has two components, (statistical) bias and variance, that are explained in layman's term here[20]. Section 3.0 and Appendix A already emphasize that the survey weight variable is to be used when producing an estimate

18. "Complex" survey data are data collected by means of a sample design featuring any combination of the following four features: stratification, clustering, unequal survey weights or finite population correction factors.
19. Data literacy training learning catalogue
20. Variance and bias

to reduce its bias to a minimum, leaving only its variance – the other component of survey error – to be assessed using the Bootstrap weight variables supplied with the SOLMP's data.

The following graph shows the distributions of weighted and unweighted estimates of the number of individuals of interest based on repeated sampling from a simulated population; it allows to compare their biases and variances.

Figure E1.1
Distributions of weighted and unweighted counts



Source: Author's image.

The unweighted estimates are clearly biased: the center of the blue distribution is located far to the left of the true count of interest indicated by the grey vertical line. This bias was to be expected since an unweighted count corresponds to the number of individuals of interest in the sample with no attempt being made to compensate for those also of interest in the population who were not sampled. In contrast, the weighted counts (which are obtained by adding up the survey weights of the individuals of interest who were selected) are unbiased: the center of the red distribution coincides with the true count, the grey vertical line.

The graph suggests that the survey weight variable converted[21] the bias in the unweighted counts into additional variance for the weighted estimates: the red distribution is wider than the blue one. Because the bias is difficult (and often impossible) to assess in practice as its magnitude depends on the (unknown) true count, it is advantageous to have all the error in variance form which can be assessed using the Bootstrap weights supplied with the data.

Some analysts are against the indiscriminate use of the survey weight variable as a means of getting valid statistical conclusions from complex survey data. They rightfully claim that there are situations where using the survey weight does not guard against bias and only works to increase the variance associated with the estimate. Hence, they argue that in some cases an unweighted estimate of a proportion would be more accurate than a duly weighted one.

21. A more accurate statement is that the survey weight variable has likely eliminated some of the sampling error initially present in the unweighted estimates and converted the rest of it into variance for the weighted estimates.

Their rationale is that when the variable of interest is independent of the sample design and collection methods used, the weighted estimate of a proportion will have no less bias than its unweighted counterpart and possibly a larger variance. While this is correct, it is usually very difficult (if not impossible) in practice to capitalize on such opportunities should they occur. For one, analysts are not expected to have the in-depth knowledge of survey methodology in general, nor how it was specifically applied to the SOLMP, to conclusively defend the assumption of independence of the analysis from the survey process. Moreover, depending on the analytical context, the survey information needed to support such a claim may not even be available in the analytical file. For these reasons, analysts are advised to use the survey weight variable in any analytical situation to protect their results from the bias induced by the survey process, even though in some situations a more accurate estimate could be obtained without using the survey weight.

E3. Two key components to a significance test

A statistical test is an advanced technique to be used as part of an analytical plan to see just how the data collected is consistent with an assumption already expressed about the population from which the sample was drawn. For example, it could be assumed that the proportion of persons in the population with a certain characteristic are the same for two groups of interest.

There are two key components to a statistical test: a statistic and its associated reference distribution. A survey's analytical file is a vast and complex array of information. The purpose of the statistic is to summarize the file's information that is deemed relevant to the analytical context into just one numerical value. For the example, a natural test to conduct involves the one-sample t statistic which is built around the difference of the estimated proportions for the two groups and its variance.

But because of survey error, even when the proportions for the two groups are equal in the population, one would not expect the statistic's value computed from the survey data to necessarily be zero, although it ought to be close to zero.

In the presence of survey error, it is natural to ask: how far must the statistic's value deviate from zero to confidently claim that the data is not consistent with the assumption of equal population proportions? This is where the reference distribution associated with the statistic, the second key component of a significance test, comes into play. It indicates which values are plausible for the test statistic given the survey error when the underlying assumption about the population is true.

E4. Significance level: drawing a line in the continuum of plausible values

As one might expect any admissible value for test's statistic is plausible (or compatible) to a certain degree with the stated assumption about the population. Still, in this continuum of plausibility one must draw the line somewhere, identifying a threshold beyond which a statistic's value is considered too implausible for the working assumption to be consistent with the data.

The threshold is normally set so that if the assumption were true – meaning that no matter how extreme the statistic value computed from the data, one would know for a fact that it is consistent with the assumption – then the analyst would accept to erroneously reject the assumption 5% of the time. That is, the analyst would know that for 5% of the samples the survey error is misleading them into rejecting a true assumption. (However, the analyst will not know in practice whether the one survey sample is one of those or not.) This probability is called the significance level of the test, and the role of the reference distribution is to turn that probability into a threshold value for the test statistic.

Complementarily, despite the uncertainty brought on by survey error, the analyst would correctly avoid rejecting[22] a true assumption (100-5)%=95% of the time; this probability is called the confidence level of the test. This notion is reminiscent of confidence intervals, and it is no coincidence since significance testing and parameter estimation

22. While this amounts in everyday language to accepting the assumption, in statistical practice we do not (fully) accept an assumption as the truth but rather say, as it is done here, that we do not find evidence to justify rejecting it.

are both rooted in the same statistical framework. In fact, we will see below how significance testing can be carried out using a confidence interval.

A statistical test is valid when its effective significance level corresponds to the targeted (or nominal) level. Hence, a test that is found to reject a given true assumption 40% of the time is not valid at the usual targeted 5% level but would be valid if 40% were indeed the targeted significance level. In the latter case, however, a statistically significant result would carry little scientific weight because with such a high false-positive rate just about any blip in the data may cause the test to reject the underlying assumption.

Remark: while 5% is a common choice, a lower significance level of 1% may be called for when the analyst is facing a greater analytical burden of proof, say to inform policy-making decisions rather than simply to figure out what avenue to explore next.

E5. Assessing significance in practice

A well-defined analytical plan may involve computing the value of a statistic using the data, which will be compared to the threshold value provided by the appropriate reference distribution and corresponding to the targeted significance level, say 5%. The associated result is deemed statistically significant at the targeted level when the computed value for the test is more extreme than the threshold value.

However, in practice, analysts commonly rely on the p-value routinely provided by statistical packages along with the computed statistic value to determine statistical significance. The p-value is the probability of getting a statistic value at least as extreme as the one just computed when the underlying assumption is true. Thus, the p-value matches the significance level when the computed statistic equals the threshold value.

In essence, the use of a p-value involves comparing two probabilities – the p-value against the targeted significance level – rather than comparing two values as before (the statistic's value computed from the data against the threshold value corresponding to the targeted level).

While p-values have always been popular, they are often misused and misinterpreted by practitioners in all areas of applied statistics, including surveys. For instance, while a low p-value does reveal something about the working assumption, it says nothing about the strength of a competing assumption considering the same data evidence. Indeed, for all the analyst knows, perhaps the only other assumption making any sense is even less consistent with the data than the working assumption being rejected.

Convenient as it may be, just looking out for results with small p-values to report on is no substitute for doing the groundwork needed to devise and execute a well-defined, relevant analytical plan. In fact, the deliberate practice of mass-producing statistical results solely to report as significant findings those with low p-values is a serious inferential issue known as p-value hacking.

E6. Specifying the reference distribution

Since the validity of a statistical conclusion hinges on the reference distribution, it is important to assign the right one to the statistic for the effective significance level to match the targeted level. In practice, there is often not just one reference distribution associated with a given statistic, but rather a whole family of those.

For example, the t-test discussed earlier is associated with the family of distributions known as the Student distributions: these distributions are analytically identical except for a parameter they share whose value determines how wide or narrow each distribution is. This parameter is commonly referred to as the number of degrees of freedom associated with the distribution (family). In the case of the Student, the larger the number of degrees of freedom, the narrower the distribution and thus the narrower the range of values it deems plausible. Hence, a narrower distribution will lead to rejection more frequently – justly or not – than a wider one would. Consequently, specifying the right parameter value – that is, identifying the correct range size for the plausible values under the targeted significance level – is paramount to getting a valid test.

However, specifying the correct number of degrees of freedom to use with complex survey data can be challenging, as various aspects of the sample design other than the sample size alone play a role. In practice, one commonly relies on the approximate rule which in the case of SOLMP equates the number of degrees of freedom to the available sample size for the subpopulation of interest minus the number of strata that are involved in the analysis. (However, because there is no way of identifying the strata from SOLMP's analytical file, it is recommended to use 1 as the number of strata involved in the analysis.)

Warning: many software packages e.g., SAS, use by default the number of Bootstrap weights supplied as the number of degrees of freedom. Since there are 1,000 Bootstrap weights for SOLMP, this would overstate the correct number of degrees of freedom to use in all analyses but those involving large domains. Thus, analyses for small to medium sized domains are at risk of making the evidence found in the data for rejecting the assumption look stronger than it really is. Analysts ought to examine the software documentation for the option of specifying their own number of degrees of freedom.

E7. Carrying out significance testing using a confidence interval

As mentioned already, significance testing is closely related to estimation using a confidence interval; in fact, the latter can be used to carry out the former. For instance, one can reject at the 5% significance level the assumption of equal proportions between two groups when the (100-5)%=95% confidence interval associated with the estimated difference in proportions does not contain 0. In essence, the confidence interval corresponds to the non-rejection area of the associated test.

Remark: this is not equivalent to rejecting when the two 95% confidence intervals associated with the estimated proportions do not overlap, which is a more conservative approach rejecting less frequently than the usual test, including when the true difference in proportions is in fact nonzero.

An alternative to using a Student confidence interval is provided by a Bootstrap Percentile confidence interval – See Appendix D for details. A distinctive feature of the Bootstrap method is that it supplies its own reference distribution based on the data. Because the reference distribution employed by the Bootstrap is not supplied by the user (either directly or because of the software's default settings), there is less of a chance of a misspecifying it.

E8. Concluding remarks

This brief exposé raised awareness to the use of the survey weight and Bootstrap weight variables, as well as to some of the most common and important issues encountered when conducting significance testing in the presence of survey data. The intent was to provide the necessary foundation for the reader to further grasp the available literature and consult with experienced data analysts or survey statisticians, ultimately leading to a successful and meaningful analytical plan. Some of the key takeaways are:

- Devise an analytical plan in collaboration with an experienced survey data analyst or a survey statistician;
- Use the survey weight and Bootstrap weight variables to deal with survey error;
- A valid statistical test requires that the effective significance level corresponds to the reported targeted level – mis-specifying the reference distribution and performing multiple comparisons are common situations prone to yielding an invalid test;
- Do not blindly rely on the p-values readily made available by software packages to draw statistical conclusions from the data.

References

Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, Statistical Science, Vol. 1, No. 1, 54-77.

Efron, B. and Tibshirani, R. J. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC.

Gagné, C., G. Roberts and L.-A. Keown. "Weighted estimation and bootstrap variance estimation for analyzing survey data: How to implement in selected software." The Research Data Centres Information and Technical Bulletin. Winter 2014. Vol. 6 No.1 Statistique Canada No 12-002-X No 2014001

Heeringa, S.G., B.T. West and P.A. Berglund. 2020. Applied Survey Data Analysis. Second Edition. CRC Press.

Korn, E.L., and Graubard, B.I. (1998). "Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data". Survey Methodology, 24, 193-201.

Kott, P.S. and Carr, D.A. (1997). "Developing an Estimation Strategy for a Pesticide Data Program." Journal of Official Statistics, Vol. 13, No. 4, 367-383.

Lewis, T.H. 2016. Complex Survey Data Analysis with SAS. CRC Press.

Liu, Y.K. and Kott, P.S. (2009). "Evaluating Alternative One-Sided Coverage Intervals for a Proportion." Journal of Official Statistics, Vol. 25, No. 4, 569-588.

Lohr, S. 2021. Sampling: Design and Analysis. Third Edition. CRC Press.

Lumley, T. 2010. Complex Surveys – A Guide to Analysis Using R. Wiley.

Neusy, E. and H. Mantel. (2016). "Confidence Intervals for Proportions Estimated from Complex Survey Data". Proceedings of the Survey Methods Section of the Statistical Society of Canada Annual Meeting, June 2016.

Pépin-Filion, D., L. Cornelissen and É. Lemyre, 2024, "Situation of English-speaking populations in Quebec and French speaking populations in Canada outside Quebec: Results of the 2022 Survey on the Official Language Minority Population." Ethnicity, Language and Immigration Thematic Series, Statistics Canada Catalogue no. 89-657-X 2024008.

Savard, Sarah-Anne (2023). Overview of the quality component of the dissemination strategy for the 2022 Canadian Survey on Disability, October 2023, Internal document.

Statistics Canada. 2020. GTAB specifications V21. Internal Document. Ottawa, Ontario.

Statistics Canada. 2023. Canadian Survey on Disability, 2022: Concepts and Methods Guide. Catalogue no. 89-654-X2023004. Ottawa, Ontario. (accessed December 12, 2023).

Statistics Canada. 2023. Evaluating confidence interval methods for the 2021 Census using the Census Monte Carlo simulation environment. Working document. Internal Document. Ottawa, Canada.

Statistics Canada. 2023. Sampling and Weighting Technical Report, Census of Population, 2021. Catalogue no. 98-306-X2021001. Ottawa, Ontario. (accessed December 12, 2023).

Statistics Canada. 2024. Canadian Survey on Disability, 2022: User Guide to the Analytical Data Files. Internal Document available in Research Data Centers. Ottawa, Ontario.

Tibshirani, Robert (1984). Bootstrap confidence intervals, Department of Statistics Stanford University, Technical report, No. 3.

Wilson, E.B. 1927. "Probable Inference, the Law of Succession, and Statistical Inference." Journal of the American Statistical Association, Vol. 22, no. 158. p. 209-212.