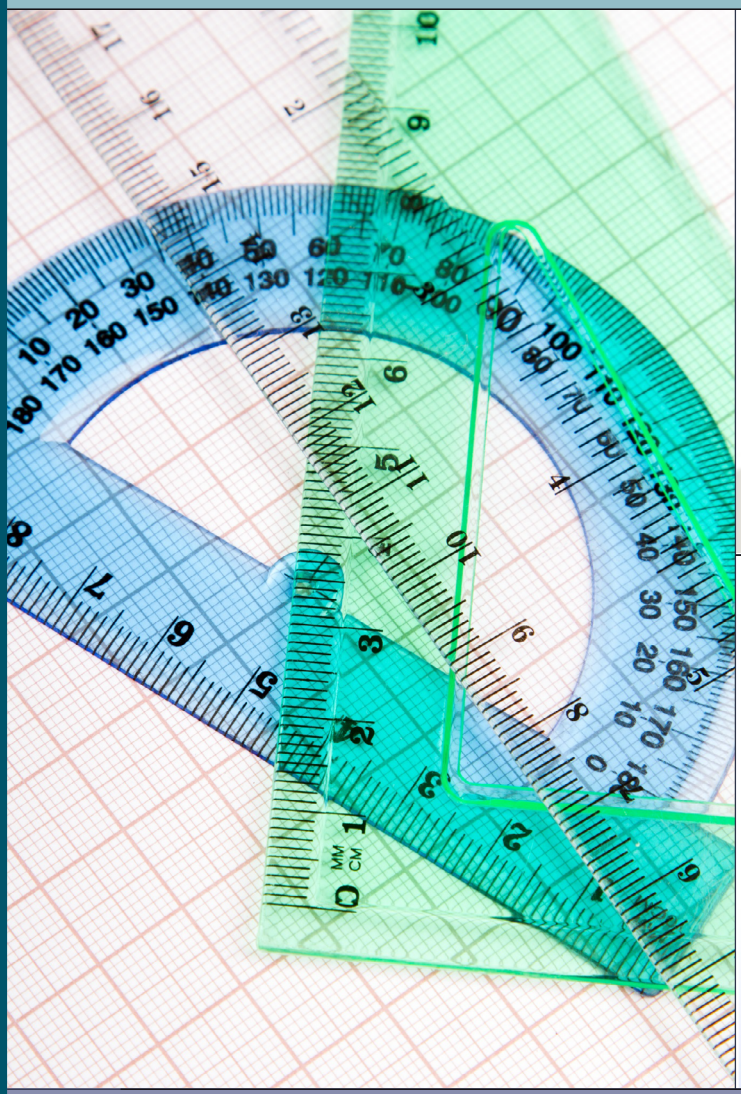
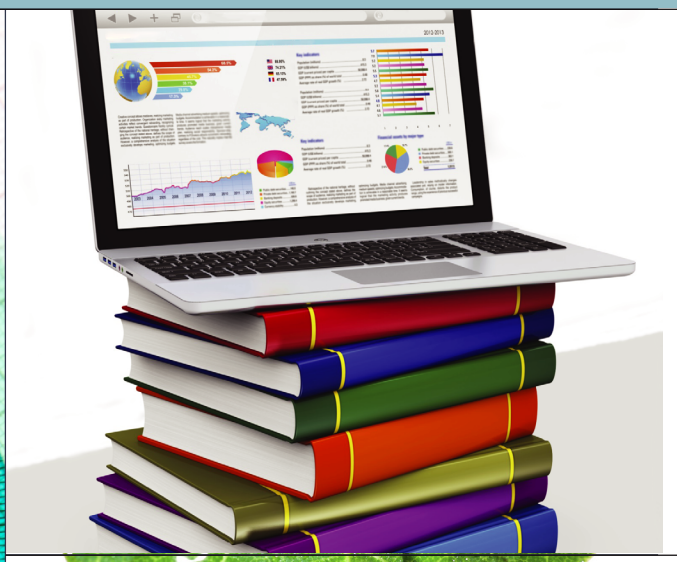
Pan-Canadian Assessment Program

# PCAP 2019

## Technical Report





cmec

Pan-Canadian Assessment Program

# PCAP 2019

## Technical Report


**Authors**

**Kathryn O'Grady**, Council of Ministers of Education, Canada
**Koffi Houme**, Council of Ministers of Education, Canada
**Yitian Tao**, Council of Ministers of Education, Canada
**Gulam Khan**, Council of Ministers of Education, Canada
**Allison Chapman-Chin**, Council of Ministers of Education, Canada



cmec
Conseil des ministres de l'Éducation (Canada)
Council of Ministers of Education, Canada

The Council of Ministers of Education, Canada (CMEC) was formed in 1967 by the provincial and territorial ministers responsible for education to provide a forum in which they could discuss matters of mutual interest, undertake educational initiatives cooperatively, and represent the interests of the provinces and territories with national educational organizations, the federal government, foreign governments, and international organizations. CMEC is the national voice for education in Canada and, through CMEC, the provinces and territories work collectively on common objectives in a broad range of activities at the early childhood, elementary, secondary, and postsecondary levels.

Through the CMEC Secretariat, the Council serves as the organization in which ministries and departments of education undertake cooperatively the activities, projects, and initiatives of particular interest to all provinces and territories. One of the activities on which they cooperate is the development and implementation of pan-Canadian testing based on contemporary research and best practices in the assessment of student achievement in core subjects.

**Note of Appreciation**

*The Council of Ministers of Education, Canada, would like to thank the students, teachers, and administrators whose participation in the Pan-Canadian Assessment Program ensured its success. The quality of your commitment has made this study possible. We are truly grateful for your contribution to a pan-Canadian understanding of educational policy and practices in mathematics, reading, and science at the Grade 8/Secondary II level.*

Council of Ministers of Education, Canada
95 St. Clair Avenue West, Suite 1106
Toronto, Ontario M4V 1N6
Telephone: 416-962-8100
Fax: 416-962-2800
E-mail: cmec@cmec.ca



*Ce rapport est également disponible en français.*

# Table of Contents

# List of Figures



# List of Tables

# 1 Introduction: What Is the Pan-Canadian Assessment Program?

## Context

Canadian ministries and departments of education have been participating in a number of assessments for approximately 20 years to measure students' skills in reasoning, problem solving, and communication to help prepare students for the future. At the international level, through the Council of Ministers of Education, Canada (CMEC), students have participated in the 2000, 2003, 2006, 2009, 2012, 2015, 2018, and 2022 Programme for International Student Assessment (PISA) (involving over 80 countries and economies in 2018), the 2011, 2016, and 2021 Progress in International Reading Literacy Study (PIRLS) (involving over 60 countries), the 2015 and 2019 Trends in International Mathematics and Science Study (TIMSS) (involving approximately 65 countries), and the 2013 International Computer and Information Literacy Study (ICILS) (involving approximately 20 countries). Most provinces/territories also conduct their own evaluations of students at different stages in their schooling. To examine the teacher context, some provinces have participated, through CMEC, in the Teacher Education and Development Study in Mathematics (TEDS-M) in 2008 and the Teaching and Learning International Survey (TALIS) in 2013. The Program for the International Assessment of Adult Competencies (PIAAC) was conducted in 2012 (and will be conducted again in 2022) as a broad study of adult literacy, numeracy, and problem solving involving 25 countries, including Canada. Canadians have long been interested in how well their education systems are meeting the needs of students and society.

## Pan-Canadian assessment

To study and report on student achievement in a Canadian context, CMEC initiated the School Achievement Indicators Program (SAIP) in 1989 to assess the achievement of 13- and 16-year-old students in Canada. SAIP was a cyclical pan-Canadian assessment program that examined student achievement in reading and writing, mathematics, and science between 1993 and 2004. In 2003, the provincial and territorial ministers of education, through CMEC, agreed to develop the Pan-Canadian Assessment Program (PCAP) to replace SAIP. The major domain of each PCAP assessment is one of these areas of learning, but each assessment includes the other two subject areas as minor domains.

School programs and curricula vary from province to province and from territory to territory across the country, so comparing results from these programs is a complex task. However, young Canadians in different provinces and territories learn many similar skills in reading, mathematics, and science. PCAP was designed to determine whether students across Canada reach similar levels of performance in these core disciplines at about the same age, and to complement existing provincial/territorial assessments with comparative Canada-wide data on the achievement levels attained by Grade 8/ Secondary II students.[1]

PCAP is designed as a system-level assessment to be used primarily by provincial ministries and departments of education to examine their education systems. Information gathered by each assessment has given ministers of education a basis for examining curricula and other aspects of school systems.

[1] PCAP is administered to students in Secondary II in Quebec and Grade 8 in the rest of Canada.

PCAP data are reported at provincial/territorial levels, by language of the school system, and by gender. The goal of national and international large-scale assessments is to provide reliable information about academic achievement and to gain a better understanding of the contextual factors influencing it. They provide policy-makers, administrators, teachers, and researchers with powerful insights into the functioning of education systems and how they might be improved. However, it should be noted that these assessments are not designed to report valid results at the student, school, or school board/district level.

In 2007, PCAP was first administered to 13-year-old students. As of 2010, it is administered to Grade 8/Secondary II students, and, whenever possible, intact classes are selected to minimize the disruption to classrooms and schools.

PCAP does not address individual student performance, nor does it involve comparisons between students, schools, or school boards/districts. PCAP results are not made available to teachers, school boards/districts, regions, or ministries/departments of education to assess students' school performance.

## Participation

Ten provinces and one territory (Yukon) in Canada participated in the first two administrations of PCAP in 2007 and 2010. Ten provinces participated in PCAP 2013, 2016, and 2019. Northwest Territories previously participated in SAIP.

## Administration time

In PCAP 2019, students were allotted 90 minutes to respond to the assessment items. They were entitled to an additional 30 minutes to complete the test, if necessary. Further additional time could be given to students for whom this type of accommodation was provided in their regular school program. After completing the cognitive test, students had 30 minutes to answer the Student Questionnaire. Students were allowed to use the resources they normally have access to in language arts, mathematics, and science classes. The Teacher Questionnaire and School Questionnaire were also administered to obtain a more holistic view of Canadian education systems.

## PCAP in both official languages

The results obtained from students educated in francophone school systems of their respective provinces are reported as "French." The results obtained from students educated in anglophone school systems of their respective provinces are reported as "English." Within anglophone school systems, although students in French immersion programs could, at the discretion of the school, complete the PCAP test in either English or French, their results were reported with those of the English-language cohort. As a resource for French-immersion students, a list of common science and mathematics terms was provided in English and French.

# 2 Design and Development of the Assessment

To avoid language bias, the PCAP assessment instrument was jointly designed in French and in English by francophone and anglophone education specialists. All items in each of the three subjects were written in both languages, and all students who took part in the PCAP field test and main study responded to the same questions, regardless of language. Samples in PCAP were selected to represent both majority and minority official-language groups[2] in the eight provinces that had sufficient numbers for valid statistical comparisons. Due to the small sample sizes, results for francophone school systems were not reported for Prince Edward Island and Newfoundland and Labrador; however, these results are included in the calculations for the overall Canadian and provincial means and totals.

## PCAP assessment cycle

PCAP assessments are administered every three years to students who are in Grade 8/Secondary II. Each assessment cycle collects achievement data using a cognitive test with a major emphasis on one of the three learning domains — reading, mathematics, and science — and a minor emphasis on the remaining domains. PCAP also collects a significant range of contextual information (e.g., on demographics, socioeconomic factors, and school teaching and learning conditions) to enhance interpretation of student performance.

Each PCAP assessment includes questions on all three domains, although the focus shifts, as shown in Table 2.1 below. The repetition of the assessments at regular intervals yields timely data that can be compared across provinces and territories, and over time. For the fifth assessment, in 2019, the focus was on mathematics, as it had been in 2010, with reading and science as the minor domains.

**Table 2.1 PCAP assessment schedule**

| Domain | Cycle 1 | | | Cycle 2 | | |
|---|---|---|---|---|---|---|
| | Spring 2007 | Spring 2010 | Spring 2013 | Spring 2016 | Spring 2019 | Spring 2023[3] |
| Major | Reading | Mathematics | Science | Reading | Mathematics | Science |
| Minor | Mathematics | Science | Reading | Mathematics | Science | Reading |
| Minor | Science | Reading | Mathematics | Science | Reading | Mathematics |

[2] With respect to the two official languages in Canada, English is the majority language outside of Quebec — across the country, 64 percent of Canadians report speaking English most often at home. In Quebec, French is the majority language — 79 percent of people in Quebec report speaking French most often (Statistics Canada, 2020).

[3] The administration of PCAP 2022 has been delayed until 2023 in response to health concerns related to the global pandemic and to minimize overlap with PISA, which has been delayed one year, from 2021 to 2022.

## Reporting PCAP achievement over time

One of the strengths of PCAP is its measurement of changes over time in student performance. PCAP achievement scales provide a common metric that provinces/territories can use to compare students' progress at the Grade 8/Secondary II level in the three core subjects from assessment to assessment. Items that were administered in the baseline years, known as "trend items" or "anchor items," will provide the basis for linking the assessment results. This basis will enable provinces/territories to have comparable achievement data from 2007, 2010, 2013, 2016, and 2019, and to plot changes in performance over time.

In 2010, there was a shift in the population definition from an age basis (13-year-olds) to a grade basis (Grade 8/Secondary II). Because the results were scaled separately on the two assessments to a mean of 500 and a standard deviation of 100, it was necessary to rescale the scaled scores from the 2007 administration to the metric of the 2010 administration. This rescaling caused variation in the 2007 means reported for reading between the two reports published in 2007 and 2010.

## Updating the assessment framework

Updating the PCAP assessment framework for 2019 began with reviewing and modifying the assessment frameworks that specify the content to be assessed. While school programs differ from one part of the country to another, PCAP is based on curriculum areas that are common to them at the Grade 8/Secondary II level. This focus on common curriculum areas allows comparisons to be made across provinces and territories of students at a comparable point in their schooling. The *PCAP 2019 Assessment Framework* (CMEC, 2020) provides the theoretical underpinnings, design principles, and performance descriptors that were used to develop test items in each of the three domains for the second cycle of PCAP (2016–23). Chapter 2 describes the major domain of mathematics, while Chapters 3 and 4 describe the minor domains of science and reading, respectively.

For 2019, the mathematics framework was updated to better reflect curricula and standards across Canada. The reading framework had been updated for PCAP 2016, when reading was the major domain for the second time, and the science framework had been updated for PCAP 2013, when science was the major domain for the first time, and so the framework content remained unchanged from those two domains. The updates were discussed and revised by provincial experts.

## Transition to PCAP online

PCAP 2019 was the first time PCAP was administered as an online assessment. To control for effects between the paper and online versions of the test and to allow data linking across the two versions or modes, a mode study was administered to a smaller proportion of students, who completed the cognitive test and the questionnaire on paper. The technical advisory group for PCAP reviewed the paper-based and online results and concluded that the results were comparable, both for PCAP 2019 and for comparisons over time. All Teacher and School Questionnaires were administered online in 2019.

## Assessment design

The PCAP assessment covers three assessment domains: mathematics, reading, and science. Each domain goes through a rotation. This is usually a nine-year period that begins with the topic as the major domain and with the development of a new or revised framework reflecting the current best thinking about assessment in that topic. The rotation continues with two subsequent PCAP assessments, in which the topic is the minor domain. The rotation concludes with the topic becoming the major domain once again and with another revision of the framework. In the 2019 assessment, mathematics was the major domain, while reading and science were the minor domains. The focus changes with each assessment, so mathematics will become a minor domain and science the major domain in the next PCAP study, in 2023 (a delay of a year from the usual rotation, due to the global pandemic).

To ensure that trends can be measured over longer periods of time, every time the framework for a major domain is revised — i.e., at the beginning of each domain rotation — a new set of items is developed to reflect the evolution of the construct. For PCAP 2019, the revised framework for mathematics and the introduction of computer-based items broadened the construct beyond what was measured in 2010, the last time that mathematics was a major domain. This means that the PCAP 2019 mathematics scale must represent the revised framework while being linked to the existing scale represented by the previous framework through the set of existing trend or anchor items.

For the PCAP assessment, eight clusters of mathematics assessment units were distributed within four test versions or booklets, so that each booklet typically contained two clusters of mathematics items, one reading cluster, and one science cluster. The four booklets were randomly and equally distributed to students within a single class. Thus, every student completed two of the eight clusters of mathematics assessment items; however, all eight clusters were completed by students within a class. In addition, pairs of booklets contained sets or units of common items allowing for comparative measurements of student performance from one booklet to another. After the cognitive assessment, students completed the Student Questionnaire.

Table 2.2 shows the distribution of the clusters, contexts (or scenarios or passages), and items for mathematics, reading, and science across the four booklets, while Table 2.3 shows the distribution of the item types for the three domains.

**TABLE 2.2 Number of clusters, scenarios, and items by domain and booklet**

| | Mathematics | | | Reading | | | Science | | |
|---|---|---|---|---|---|---|---|---|---|
| | Clusters | Scenarios | Items | Clusters | Passages | Items | Clusters | Scenarios | Items |
| **Booklet 1** | 2 | 7 | 22 | 1 | 2 | 10 | 1 | 3 | 9 |
| **Booklet 2** | 2 | 7 | 19 | 1 | 3 | 12 | 1 | 3 | 8 |
| **Booklet 3** | 2 | 7 | 15 | 1 | 3 | 9 | 1 | 3 | 8 |
| **Booklet 4** | 2 | 6 | 17 | 1 | 3 | 9 | 1 | 3 | 8 |

**TABLE 2.3 Distribution of items by type and assessment domain**

| Domain | Booklet 1 | | Booklet 2 | | Booklet 3 | | Booklet 4 | |
|---|---|---|---|---|---|---|---|---|
| | SR* | CR** | SR | CR | SR | CR | SR | CR |
| **Mathematics** | 17 | 5 | 14 | 5 | 13 | 6 | 14 | 6 |
| **Reading** | 10 | 2 | 12 | 2 | 7 | 2 | 7 | 2 |
| **Science** | 7 | 2 | 7 | 3 | 8 | 2 | 11 | 1 |
| **Total** | 34 | 9 | 33 | 10 | 28 | 10 | 32 | 9 |

* Selected response
**Constructed response

## Working groups

The primary focus of PCAP 2019 was mathematics. Thus, new cognitive items were developed for that domain only, and the questionnaires were revised to focus on the teaching and learning of mathematics. Working groups consisted of experts in mathematics curriculum, as well as in teaching, learning, and assessment. They came from various provinces, and almost half of the participants were bilingual. These experts were extensively involved in PCAP and took part in various stages of the project, such as developing the assessment framework, drafting items, validating and editing items, and comparing English and French items. Some also participated in coding sessions for the field test and main study.

## Updating the mathematics framework

This framework was reviewed and revised by mathematics experts to reflect current thinking in the field of mathematics literacy and numeracy research, including the assessment of mathematics.

The mathematics framework review working group comprised mathematics curriculum and assessment specialists who reviewed and further revised the framework to reflect changes in the mathematics program of studies in provinces across Canada.

## Item development

Documentation to guide all stages in the item-development process was prepared for the meetings of test developers that took place in Ottawa in July 2017 and in Toronto in August 2017.

The orientation included an overview of the mathematics assessment framework, the development process and timelines, specification of item requirements, and the importance of framework fit. The session began with a large group discussion to identify topics that would be of interest to Grade 8/ Secondary II students and that would fit within all programs of study in Canada for this age group. Item development took place in small groups and happened simultaneously in English and French. The sessions involved an iterative process: small groups worked to develop a unit that contained a series of questions around a stimulus that had a good fit to the framework. A complete unit consisted of the stimulus material, four to seven items with a mix of both selected and constructed response types, and a guide to coding the responses to each question. Each coding guide was made up of a list of response categories (full, partial, and no credit), each with its own coding code, descriptions

of the kinds of responses to be assigned each code, and sample responses for each response category. The units were presented to the large group for discussion regarding item quality, age appropriateness, cultural and gender sensitivity, curriculum coverage, and framework fit. Following the discussion, the small groups revised their items by incorporating the suggestions and recommendations from the large group. Upon completion of the first round of unit development, the large group reassembled to choose their next stimulus topic, which helped to ensure a broad coverage of the mathematics framework.

At the conclusion of the item-development sessions, the working group reviewed and revised the mathematics framework so that it adequately reflected the topics and types of questions that could represent the commonalities among the Grade 8/Secondary II programs of study in Canada. Small groups then reviewed mathematics items from previous administrations to ensure that they adequately represented the framework and were properly classified.

The completed items, which were developed in both English and French, were translated and copy-edited at the CMEC Secretariat.

## Item review

Three field test booklets were designed to include approximately the same number of questions for each of the four subdomains in mathematics (numbers and operations, geometry and measurement, patterns and relationships, and data management and probability), and the same number of questions at each cognitive level and of each item type (i.e., constructed and selected response). The booklets were reviewed by provincial experts in mathematics. The items were reviewed for content, vocabulary, translation, program of studies fit, and freedom from bias, and to verify the classification of the items for the subdomains delineated in the framework.

At the end of this review, the framework was again reviewed and revised to better reflect the common elements of the program of study documents from the provinces. In some cases, items were removed because of biases with respect to gender or culture or because the items were problematic after translation. The remaining units were edited and verified in both languages.

## Verification of the assessment items and coding guides

Before including items in an assessment, whether for the field trial or the main PCAP administration, it was important that these items be reviewed from various perspectives by groups of experts to ensure that items were sound and would provide an accurate assessment of the skills of Grade 8/Secondary II students across the country. The validation process, which was done by groups of experts, included the following steps:

- translating and comparing items in English and French
- ensuring that items were equivalent in both languages with respect to difficulty
- verifying the classification of mathematics items based on the mathematics framework
- verifying the coding guides for constructed-response items
- editing for language and style, technical editing, and psychometric editing

## Editing for language and style

An important step in the review of items is editing for language and style. The language editing had to address grammar, syntax, spelling, and punctuation for each item, scenario, or graphic in each assessment booklet. The stylistic editing then had to check spaces, fonts, number of lines, page composition, and the introduction to each statement. Editors had to verify that font size was the same for all items; spaces between lines in an item were the same throughout the booklets; page composition was consistent between the online and paper-based booklets; each item began with a statement followed by a question; the amount of space provided for the student's answer was appropriate for the length of the expected answer; and sources were accurate, which means that when an item referred to a graphic on another page, the reference was in fact to the correct page.

## Technical editing

Technical editing checks and validates the correct answers, calculations, data, etc. The four versions of the test contained selected-response items with four possible answers. Editors had to ensure and verify that there was in fact only one correct answer and that the three other choices were logical distractors. In science and mathematics, an item could require students to perform a calculation to obtain the correct answer. The calculation therefore had to be repeated to ensure that the final answer was one of the selected-response answers. Although there were no selected-response answers to check for the open-response items, the items (and sample answers) still had to be validated again to ensure that the correct descriptors were assigned and checked for accuracy, either by referring back to the text or performing the calculations.

Several mathematics and science questions or scenarios included tables, diagrams, and charts with data. Editors therefore had to verify and ensure the accuracy of the information. Students might also have to refer to a table or chart to obtain a correct answer. In the item, students were told on which page the table or chart in question could be found. Editors therefore had to ensure that the page number the students were directed to was correct.

Several reading questions had line or paragraph numbers. Editors verified that the numbering system was consistent between versions of the test. In the case of anchor items, it was also verified that the items were identical in booklets from different PCAP administrations.

During item editing, it was important to verify that all components of a text or item were present so that students would be able to answer the question. If, for example, components were missing from the item, students would be unable to answer the question correctly, and these items would have to be excluded from the analysis. It would be unfortunate to have to remove an item from the test, especially if that item could have been useful in measuring students' skills.

## Psychometric editing

The experts in mathematics, reading, and science conducted a psychometric edit of items. For selected-response items, one factor to be checked was the order of the possible answers. When the possible answers were text, the distractors were ordered from the shortest sentence or word to the longest. When the possible answers were numbers, the distractors could be placed in increasing order, from the smallest to the largest. This approach to ordering possible answers thus placed the correct answer in random order. Each possible answer also had to be approximately the same length.

If one choice was more detailed, students would be more inclined to opt for this choice. It was also important to check the accuracy of the correct answers to ensure that there was not a second answer that might also be correct, to avoid any ambiguity.

A coding guide with descriptors was developed for new constructed-response items by experts in mathematics. The coding guides for trend items remained unchanged from the previous administration in which each subject was the major domain (2007 and 2016 for reading, 2010 for mathematics, and 2013 for science) to ensure consistent marking of items to be used in analyzing achievement changes over time. Various codes were assigned to students' answers. The codes ranged from 0 to 3. Each code included a complete description as well as one or more examples taken from students' answers. The experts therefore had to review all the coding criteria and ensure that the codes established were clear and precise. This step was very important because in the item-coding session for the three subjects, coders received training on each item to be coded. They had to be able to properly distinguish each code so they could assign the one most consistent with the student's answer.

The experts also had to review the table of specifications, which presents the master assessment plan, and validate the item types. For example, the assessment had to include a balanced mix of constructed-response items and selected-response items to make efficient use of the students' assessment time while gathering critical and personal reactions in an open context.

## Item approval by the provinces

Before including items in the field test and main study, the items selected had to be approved by the provinces. CMEC produced three field test booklets and four main study booklets, in English and French, and in paper and online formats. These were sent to the provinces for their review. CMEC obtained approval from each province to include the scenarios or passages and items in the field test in 2018 and the main study in 2019.

# 3 Development of the Contextual Questionnaires

Students participating in PCAP, in addition to their teachers and school principals, complete questionnaires that are designed to provide provinces and territories with contextual information to aid in the interpretation of the performance results. Researchers, policy-makers, and practitioners can use the information provided by these questionnaires to help them determine what factors influence learning outcomes. The content of the contextual questionnaires changes depending on which of the three domains is the primary focus of the PCAP assessment. Because the primary domain of the 2019 PCAP assessment was mathematics, contextual questions addressed factors that have been found in past studies to correlate with mathematics achievement. Some examples of these correlates include parental level of education, language spoken in the home, and the number of books in the home.

Contextual questionnaires completed by teachers cover questions about teaching and learning conditions, including teachers' homework expectations, assessment practices, areas of specialization, and years of teaching experience. The School Questionnaire, completed by the principal, is the key source of information about various dimensions of each school, including the structure and organization of the school, school climate, school policies and practices, and curriculum and instruction.

The PCAP questionnaires are available on the CMEC website, at https://cmec.ca/697/PCAP_2019.html. Access to the PCAP data set is available upon request. More details about the conceptual framework of the questionnaire component of PCAP are described in Chapter 5 of the *PCAP 2019 Assessment Framework* (CMEC, 2020). The conceptual framework reflects current research findings and best practices in the field of mathematics learning.

## Updating the questionnaire framework

Mathematics experts were asked to independently review and revise the questionnaire framework used for PCAP 2010, the previous year in which mathematics was the focus of the assessment. The experts were asked to identify important aspects of the context in which students learn mathematics, as well as to identify areas that could be improved or topics that could be added based on the academic literature on mathematics, which contextualizes how students learn mathematics and the challenges they face.

The goal of the review was to develop three concise questionnaires that focused on issues related to learning and teaching mathematics — i.e., the major domain — and that could provide important contextual information for the provinces and territories. The mathematics-focused questionnaires developed by the experts were translated and copy-edited by CMEC and sent to the members of the Pan-Canadian and International Assessments Committee for review and further revision.

There were three questionnaires included in the PCAP 2019 assessment: one for participating students, one for their Grade 8/Secondary II mathematics teachers, and one for school principals. These questionnaires also focused on the particular need to capture factors associated with mathematics achievement and were intended to contextualize the assessment results. They include some core descriptive data useful for both policy and research — for example, student socioeconomic

status (SES), school demographics, and teacher qualifications. Various topics also addressed policy-relevant issues. The questions focused primarily on the assessment's major domain, mathematics, but also included probes into teaching and learning strategies and behaviours. Other questions were in areas that support the directions identified by ministries and departments of education, even if these do not have obvious links to achievement in the major domain. This selection of topics aimed to provide information that would be useful in research applicable to mathematics.

## Gender identity

Inclusive education is valued in Canadian provinces and territories and has led to the development of policies and resources to support inclusion. One aspect of inclusive education is gender identity. In the PCAP 2016 and 2019 Student, Teacher, and School Questionnaires, the gender question was expanded to allow two additional choices for respondents, as shown below.

| How do you identify yourself? |
| --- |
| o Male |
| o Female |
| o I identify myself in another way. |
| o I prefer not to say. |

## Gender differences

Differences in mathematics achievement favouring boys have been a consistent feature of large-scale assessments, both nationally and internationally. The concern in the questionnaires was to uncover some potential explanations for gender differences by focusing explicitly on:

- differential treatment of boys and girls in school; and
- differential mathematics-related behaviours or interests outside of school.

Gender differences are also persistent in reading, although they favour girls. While this issue is less strongly emphasized for science, there remains an interest in following trends in gender differences over time across all three domains.

## Confidentiality

Both the Teacher and School Questionnaires were linked to student results but used unique identifiers to preserve confidentiality.

# 4 Sampling Procedures

In the spring of 2019, the fifth Pan-Canadian Assessment Program was administered. It assessed three domains: mathematics, reading, and science, with mathematics being the primary domain. Four assessment booklets were used in which all three domains were assessed, with the majority of the items focusing on mathematics. One school grade — Grade 8/Secondary II — was assessed. Eighteen populations[4] were involved in the assessment.

This chapter describes the assessment's sampling plan and explains how activities relating to the selection of samples took place.

## Sampling design

Between 1993 and 2004, CMEC became involved with pan-Canadian assessments through SAIP. In 2007, PCAP replaced SAIP. Although PCAP has retained some of the characteristics of the SAIP assessment, some of the technical aspects have been modified: three domains are now assessed in each cycle, one being considered the primary domain and the other two regarded as minor ones. Several assessment booklets are used. From 2010 on, the population to be assessed was defined in relation to a level of education rather than age. The collected achievement data are mainly used in two ways: to calculate performance levels for the major domain and to compile mean results for all the assessed domains. The sampling design had to be adapted to ensure data collected will serve the analytical needs and to be generalizable at the pan-Canadian level.

One major criterion for the sampling of the main study, aside from the representation of the Canadian population, is the consistency of procedures across cycles. For the main study of PCAP 2019, as was the case for the 2016 assessment, CMEC has contracted and worked closely with Statistics Canada on the sampling design, the implementation of sampling, and the assigning of weights to the data collected. This chapter provides a summary of the sampling design for PCAP 2019; greater detail on sampling and weighting can be found in the report prepared by Statistics Canada (see Appendix A).

Defining the population from which a sample is selected is an essential step in developing a sound sample design. A good definition of the target population[5] facilitates the sampling process and prevents ambiguities. Table A.3 in Appendix A presents the population represented in the PCAP sampling design. To validate the accuracy of the sampling frame, Statistics Canada compared this frame with the Canadian census population projection of 13-year-olds, the age represented by the majority of students in Grade 8/Secondary II. It was concluded that the frame aligned well with the population projection by Statistics Canada. These statistics are derived from data that the provinces supplied to CMEC for the 2018 field test of the assessment.

[4] Here, the word *population* refers to all eligible Grade 8/Secondary II students within a province and/or a linguistic group.

[5] *Target population* means the schools eligible for selection, after exclusion of the schools that don't meet the criteria adopted by CMEC or by the provinces/territories concerned. The *overall population* consists of all the schools that have Grade 8/Secondary II students.

## Sampling frames

### *First stage survey frame – list of in-scope schools*

As was the case for the SAIP assessments and previous PCAP cycles, a two-stage sampling procedure was followed for PCAP 2019. First, participating schools were selected (first stage survey frame – list of in-scope schools), and second, a Grade 8/Secondary II class was chosen in the selected schools (second stage survey frame – list of in-scope classes). Given the size of the populations being assessed, a census was taken of certain target groups' schools, and students could then be selected in those schools. In some cases, there was a census of students in Grade 8/Secondary II. The statistics produced for students in a sample had to be generalizable; therefore, each sample had to meet certain criteria.[6] These criteria concerned, in particular, the size of the sample, the a priori exclusion and inclusion of schools, and the process employed to make the selections.

In preparing the sampling frame, some schools were excluded based on the following a priori categories:

- special schools in which all students had special education needs,
- schools within another province,
- geographically isolated schools,
- federal/international schools,
- schools that are not funded, and
- schools that are closed.

It should be noted that in the first stage survey frame, exempted schools were excluded and were not considered during the sampling process.

### *Second stage survey frame – list of in-scope classes*

At the second stage of sampling, the list of in-scope classes was obtained. This frame was developed upon receiving all the class lists in the selected schools. In this stage, students were excluded based on the a priori categories below. A detailed description of these categories is available in Chapter 6 and Appendix A. In brief, the exemption criteria are as follows:

- functional disabilities
- intellectual disabilities or socio-emotional conditions
- limited language abilities in English or French (non-native speakers)

These students, or classes of students, were removed from the frame before sampling. An entire class can be exempted if all the students are in a category for which we exempt students. All exemptions at the class level had to be approved by CMEC.

[6] In the case of a census of students, there is no statistical inference, and margins of error don't usually have to be compiled.

## Stratification

*Stratification* is a means of organizing the sampling frame so that better precision can be achieved with a fixed sample size. Stratification can also be used to guarantee that a minimum sample size for certain population groups will be obtained. Strata are exhaustive and are mutually exclusive groups of schools, with each school assigned to only one stratum. The total sample size is separated among the strata, and each stratum is sampled independently.

In order to publish reliable statistics at the pan-Canadian and provincial level, as well as by the language of the school boards/districts in provinces, a large enough sample within these domains was needed. Thus, the PCAP strata are defined as the cross-classification of the province by language of the school board or school district.

In some provinces, in order for the PCAP results to be representative of both the province and language (population), a census of schools and/or census of students were used for some strata. The census of schools included all schools with Grade 8/Secondary II students as sampled schools, and the random selection of classes took place within these schools. A census of students comprised the selection of all Grade 8/Secondary II students within a selected school. Table A.5 in Appendix A outlines the list of strata and whether schools were sampled at stage one, or whether censuses were used at the school level.

## Sample sizes

Sample size is tied to the numerical size of the population, the margin of error, and the confidence level that is acceptable when statistical compilations are done so that the data can be generalized for the assessed populations.

The use of several assessment booklets and the grouping of students by performance levels have a direct impact on the size of the samples. Taking these two parameters into account, the margins of error would have considerable variations. Therefore, a sufficiently large number of students was selected to guarantee a margin of error of no more than 3 percent overall, with a confidence level of 95 percent, which was consistent with previous PCAP administrations. The formula used to determine the size of a sample in relation to the calculation of frequency distributions is shown below.

$$n = \frac{Nz^2pq}{Nd^2+z^2pq}$$

Where

$N$ = size of the population

$z$ = X-axis value on the normal curve corresponding to the desired confidence level

$p$ = proportion observed in the sample

$q = 1 - p$

$d$ = desired precision, i.e., the margin of error that is acceptable

# First stage sampling – sampling of schools

## *Databases on the schools*

To carry out the sampling work, CMEC needed to prepare a database for each population assessed. Each province had to use the same file prepared by CMEC to draw up its list of schools and prepare other necessary information. The variables required for each school include the total number of students in each school, the number of Grade 8/Secondary II students, the language of the school board/district, whether the school was an immersion school, and other school information.

## *Selection of schools*

The selection of schools was carried out by Statistics Canada, and two methods were used in this stage: the use of censuses as discussed above, and the use of Systematic Sampling (SYS) for the strata without censuses. SYS is discussed in greater detail on the sampling document in Appendix A. In this stage, both sampled and replacement schools are selected.

## *Exclusion of schools*

The decision to exclude some categories of schools, or some particular schools, was made by each provincial/territorial coordinator. However, the number of students affected by these exclusions could not exceed a certain proportion (around 2 percent) of the total population. The schools excluded from the sampling would still appear in the data files for a population that was assessed.

CMEC collected statistical information on the schools of each population using the parameters contained in the files on schools that the provinces prepared. This information included:

- the number of schools and students in the total population;
- the number of schools and students excluded from the total population;
- the number of schools and students that was part of the target population (i.e., the total population less the exclusions);
- after the selection of schools, the number of schools and students that was part of the selected sample.

If the data indicated that the exclusion criteria had not been followed (2 percent or less of students excluded a priori), CMEC contacted the provinces concerned.

It was very important that the proportion of students affected by the exclusion of certain schools complied with the established criteria. There might be a number of reasons to justify the a priori exclusion of certain schools: size, distance, special clientele, or being under the authority of a province other than the province where they are located. Coordinators had to provide CMEC with the identification numbers of the schools to be excluded and the reasons for this decision.

This information was codified in the stratum provided for this purpose. All the schools had to be included in the data files on each population assessed, since it was necessary to know, for each of these populations, the total number of students in Grade 8/Secondary II.

## Second stage sampling – sampling of students

As indicated earlier, sampling for the PCAP assessment took place in two stages. First, in cases where there was not a census of schools to participate, schools were selected. However, not all students in Grade 8/Secondary II in a selected school needed to write the PCAP assessment. Statistics Canada had to take a sample of the students who would participate. Their selection had to comply with strict rules so that the student sample would be representative of the populations being assessed. CMEC randomly chose the Grade 8/Secondary II class of the selected schools that would participate in the assessment. The following process was used to select students:

1. First, each provincial coordinator submitted a list of all eligible schools with Grade 8/Secondary II students that were under the respective province's authority.

2. CMEC selected schools to participate in PCAP and sent the List of Schools to provincial coordinators.

3. The coordinators contacted the selected schools and asked for a list of Grade 8/Secondary II classes. This list was submitted to CMEC.

4. CMEC selected classes to participate in PCAP and sent the List of Classes to provincial coordinators. It was possible that, in some cases, more than one class was chosen in the same school. After consultation with schools, provincial coordinators could decide to withdraw a class from participation. In this case, they had to communicate with CMEC so that a replacement class could be selected. Provincial coordinators had to be aware that such a replacement was allowed under only exceptional circumstances and had to be approved by CMEC.

5. The coordinators asked the selected schools to complete a List of Students for each Grade 8/Secondary II class selected to participate. The lists also indicated the names of the students who could not take part in PCAP and identified any special needs. The school principals were asked to list all Grade 8/Secondary II students as follows:

    i. When possible, a list of all Grade 8/Secondary II class groupings (e.g., 8A, 8B) that took place in the first period of the first day of the school's regular cycle (e.g., a five-day or seven-day cycle). This was Option A.

    ii. If the process in Option A was not possible, then a list of students currently registered in Grade 8/Secondary II in alphabetical order.

6. After the assessment was administered, provincial coordinators sent CMEC a list of students who participated in PCAP 2019. The same lists prepared for step 5 were used, with reasons given for any student's non-participation in the assessment.

**TABLE 4.1 Summary of the criteria used to sample students, based on the types of strata**

| Type of Stratum | Sampling Methods |
|---|---|
| Stratum sampled at two levels | (a) Schools with >20 students: Simple Random Sampling (SRS) – one Grade 8/Secondary II class; approximately one-quarter of students will be assessed using one of the four booklets<br>(b) Schools with >20 students, but unable to enumerate Grade 8/Secondary II classes: enumeration of all Grade 8/Secondary II students will be completed and an SRS of twenty students will be selected |
| Census of schools | (c) Schools with 20 or fewer students: census of students<br>One Grade 8/Secondary II class is sampled for every school on the list; one-quarter of students from each class will be assessed using one of the four booklets |
| Census of students | All the students on the list are sampled; one-quarter will be assessed using one of the four booklets |

The sampling process is a very important aspect of assessment activities such as PCAP. The credibility of the results that are made public at the end of the project often depends on it. The selection of the schools invited to participate in PCAP is made centrally on the basis of information provided by provincial coordinators. When it comes to students to be assessed, CMEC selects the class(es) in each school that is part of the chosen samples.

## Weighting of sample

Upon completion of data collection, data files with participation information were sent to Statistics Canada to compute the weights. Weights were assigned to students, teachers, and schools. Details of the weighting procedures can be found in Appendix A.

# 5 Field Testing

Items to be administered to students in large-scale performance assessments must be checked first for quality — both intrinsic quality and their appropriateness for the target population. Items developed by content experts are tested at this stage of the process. Field testing involves a larger number of items than the actual administration so that only the best items are used to assess the performance of Grade 8/Secondary II students.

## Item review

Two content experts in mathematics reviewed all mathematics items for content and verified the keys and coding guides and the items' classification for the test's subdomains, as outlined in the mathematics assessment framework. The experts also reviewed the items to identify any issues with the vocabulary level and for biases (e.g., related to gender, culture, geography). Items selected for the field test represented a broad coverage of the mathematics subdomains and a range of difficulties.

## Assessment booklets

Following the final review, the newly developed item blocks and existing trend item blocks were arranged into digital block combinations, for both PCAP online and the paper-based test. Three versions of the test, or booklets, were compiled for the field test. Each booklet followed the specifications outlined in the *PCAP 2019 Assessment Framework* and contained about 40 mathematics items in addition to the Student Questionnaire. The field test booklets were sent to provinces for review by their curriculum and assessment experts. Provinces were also asked to verify that the items were equivalent in English and French. Following the recommended revisions, the booklets were input into the online Central Coding System (CCS).

The students had 90 minutes to complete the mathematics questions in the booklet and 30 minutes for the Student Questionnaire. The Teacher and School Questionnaires were prepared as separate booklets.

## Coding session

The coding session took place over five days in Ottawa in July 2019. There were some 2,000 booklets scored, with approximately 1,000 booklets in English and 1,000 in French. There were two table leaders (one for the English table and one for the French table) and 20 coders, half of whom were assigned to the English table and the other half to the French table. Approximately two-thirds of the coders were nominated by provincial ministries/departments of education.

The coding process included twice-daily reliability cross-testing to ensure that coders evaluated items consistently and in accordance with the codes assigned by the experts. The degree of consistency between coders and experts was generally above 85 percent. For the few items that had lower consistency in the reliability review, coders reviewed the training materials and then rescored the items.

## Data capture

Students recorded their selected-response answers (e.g., multiple choice, true or false) using radio buttons or drop-down menus in the digital assessment or on a tear-out answer sheet in the paper assessment. Students wrote their answers for constructed-response questions in the space provided in the digital assessment or directly in the field-test booklets. Selected-response items were given one point per correct response. Constructed-response items, which could be awarded full or partial credit, were ranked on a scale from 0 to 3.

All student responses were scored using the online coding system. Paper responses were scanned and input into the system. Coders at the field-test coding session coded only the constructed-response items, by selecting the code for each item that best matched the student's response. Items were randomly assigned to coders. Teacher and School Questionnaires were administered online. Collected data from the assessment and all questionnaires were merged in an Excel file to create a database.

## Data analysis

Field-test data for the mathematics items were analyzed by a CMEC expert on psychometrics. The data were presented to the PCAP Technical Advisory Committee, who reviewed the analysis, databases, files, and rules for data capture (e.g., weighting of items).

Data analysis was performed using classical theory. The PCAP Technical Advisory Committee used the resulting data to identify, from a statistical perspective, the best items for the main study and any aberrant items, that is, those that did not behave like the other test items. Statistical indices were used for item analysis, including a difficulty index and a discrimination index, to check the psychometric qualities of each item. The difficulty index is based on the $p$ value, $p$ being the proportion of individuals who successfully answered the item over the total number of individuals who answered the item. Experts also verified item discrimination to ensure that each item differentiated between stronger and weaker students. The Cronbach's alpha coefficient was used to estimate internal test consistency.

Statistical experts also performed other potentially relevant analyses, such as calculating averages for each item and preparing frequency distributions for the percentage of students who selected each answer for selected-response items or who were assigned each code for constructed-response items. They also analyzed the percentage of missing data and performed differential item functioning (DIF) analysis based on language and gender.

## Item selection

Following field testing, the item-selection working group, with representatives from six provinces, met to review and select passages and items for the main administration. The group was provided with all assessment booklets, as well as results and statistics for each item, to verify item quality, degree of difficulty, and equivalent functioning in both official languages and for all genders. The item-selection process also took into account comments from coders at the field-test-coding session (from the questionnaire administered at the end of the coding session), which included some pertinent remarks on the assessment instrument in general as well as comments on each item regarding its quality.

The working group selected items for all three domains: mathematics, reading, and science. All mathematics, reading, and science items that were to be used as anchor items were identical to those used in the cycle in which these subjects were the primary focus of the assessment: PCAP 2007 and 2016 for reading, PCAP 2010 for mathematics, and PCAP 2013 for science. Because only a small subset of items was required, the working group took care to represent each subdomain and a range of difficulty levels.

The working group included bilingual experts who were also tasked with comparing the English and French versions of the booklets to determine whether students performed better on an item in one language than in the other. In the event that items did perform differently, the working group was asked about possible reasons.

The questionnaire development working group, which prepared the three questionnaires for the field test, was reconvened. The working group reviewed Student, School, and Teacher Questionnaire responses, both for content and from a statistical and psychometric point of view, and selected those items that were expected to yield the most relevant information during the main study, such as linking context data and student performance.

## Review of the assessment framework

Field testing of items yielded information that facilitated the selection of the best items for the main study. Subsequently the PCAP 2019 Assessment Framework was reviewed to ensure alignment between the framework and assessment items. Very few changes to the framework were required.

# 6 Main Study

The PCAP assessment was conducted in spring 2019, with the primary domain being mathematics. (The minor domains were reading and science.) Almost 30,000 students selected at random from over 1,500 Canadian schools in 10 provinces took part in the test in English or French. The main items assessed the knowledge and skills of Grade 8/Secondary II students in all three subject areas. The majority of students completed the assessment online; however, a mode study was conducted, with a sample of students assigned the paper-based test.

## Assessment booklets

Each assessment booklet included two clusters of mathematics items, and one cluster each of reading and science items. In order to assess the equivalency of each booklet, a subset of items for each domain was repeated in pairs of booklets. During the assessment booklets' layout, CMEC took care to include scenarios and items selected by the working groups and to display them in the same manner in both languages.

## Reviewing the assessment material

Before finalizing the assessment materials, all the provincial coordinators reviewed them so that comments could be incorporated as required. The materials sent to the provinces for review included all versions of the assessment booklets, the Student Questionnaire, the Teacher Questionnaire, the School Questionnaire, and the administrative documents. CMEC received approval of the assessment materials from each province.

## Paper-based assessments

Sample assessment booklets were reviewed to ensure that all changes made to the content by the working groups, as well as by CMEC Secretariat staff, had been incorporated into the new versions and that the paper-based and computer-based versions of the test were the same. Once this process was complete, proofs provided by the printer were reviewed and approved, after which the assessment booklets were converted to PDF format and printed. A unique identification number with a bar code was printed on the cover of each booklet so that it could be assigned to the right student. The assessment booklets and administrative documents were then packaged for each school and, based on the instructions from each province, sent either to the provincial coordinators for distribution to the schools or directly to the school coordinator.

## Online assessments

A detailed description of the online assessment booklets and test delivery system can be found in Chapter 7.

## Administrative documents

Administrative documents were printed and sent to each participating school. Each school coordinator was asked to ensure that they had the materials for their school (including assessment booklets for schools selected for paper-based tests). Any missing documents had to be reported immediately to CMEC to ensure their arrival before the scheduled test date. If the school principals and school districts/boards/commissions had any questions or needed more information about the assessment or assessment materials, they were asked to contact the provincial coordinator directly.

## Letter sent to parents/guardians of students

Prior to administering the assessment, the school coordinator had to inform the participating students as well as their parents/guardians. A brochure was distributed to parents to inform them about the assessment's intent and importance.

## Administration procedures

Each school selected had to appoint a school coordinator to administer PCAP in that school. The assessment then had to be administered according to the procedures that CMEC established to ensure that PCAP was administered uniformly in all the selected schools. Before proceeding with the assessment, the school coordinator had to become familiar with the administrative documents, in particular the *Handbook for Schools*, which outlined the test's administrative procedures. If the school coordinator had any questions related to the assessment, they had to communicate with the provincial coordinator.

Each student had a unique identification number (ID). The IDs were assigned to protect students' confidentiality. Students' names from the Student Tracking Form were used to facilitate the administration process in schools. Extra Student Login Forms and, where applicable, paper tests were provided to allow participation of new students who enrolled after the original List of Students was submitted.

Students with special needs were identified on the List of Students. CMEC provides schools with the assessment materials needed so that these students could participate in the assessment without risk of compromising its integrity. For example, large-print test formats were available in both the paper-based and online versions of the assessment to accommodate the needs of these students.

If a selected student could not participate in the assessment for any reason, the school coordinator was not allowed under any circumstances to replace that student but instead had to exempt the selected student from the assessment and indicate this on the Student Tracking Form.

## Assessment site

The school coordinators had to find a site to administer the PCAP test. It was essential to choose a quiet place where the students had access to computers and enough workspace to be able to respond to the assessment items without interruption. Wherever possible, they were advised to administer the assessment in the morning to obtain the students' best performance.

## Administering the assessment

At the start of the assessment, the school coordinators handed out one copy of the assessment booklet or a page with student login information to each student on the Student Tracking Form. The four booklets were equally distributed among the students in the class. The coordinators also had to ensure that they gave the students instructions before proceeding with the administration. They told the students that they had 90 minutes to respond to the assessment items. If necessary, the students could take 30 additional minutes to complete the assessment. They also had 30 minutes to complete the Student Questionnaire.

For each student, the school coordinators had to indicate a student participation code on the Student Tracking Form. This procedure allowed the list of selected students to be checked against the assessment booklets to determine the student's status, and whether they had participated in, had been exempted from, or had been absent from the assessment.

Once the test was completed, the coordinators collected all assessment documents and stored them in a secure place to keep the material confidential. Borrowed laptops were repacked and shipped back to CMEC or to another school in the same province.

Along with the school coordinators, provincial coordinators played a key role in ensuring the assessment's smooth administration. Provincial coordinators were responsible for observing the assessment's administration in between 5 and 10 percent of the schools in their region. They had to conduct telephone follow-up and direct observation in schools to gather the necessary information on the test's administration. If they travelled to schools for observation, they simply had to note the extent to which the correct procedures were followed. Under no circumstances could they intervene during the course of the assessment. The main elements to be observed were the security surrounding the assessment materials, compliance with the directives given to schools, compliance with the allotted time, and compliance with the rules on how to answer students' questions. These coordinators had to document their observations using the Provincial Coordinator's Report.

## Students with special needs

For this evaluation, accommodations were defined as modifications that do not compromise the integrity or content of the test, but provide an equal opportunity to all students to demonstrate their knowledge and skills at the time of the evaluation. Students requiring accommodations should have been previously identified when the school submitted its list of eligible students. The school coordinators had to notify the provincial coordinators when a student was identified as having special needs, to guarantee that special test versions were included in the shipment of assessment booklets to the school or that the accommodation was activated for the student when they logged into the assessment. It was important to make the necessary arrangements to allow students with special needs to participate in the assessment as much as possible without compromising the assessment's integrity.

Accommodations were permitted only for those students who normally benefit from them during their regular classroom work. Authorized accommodations included Braille (paper-format tests only), large print, coloured background, and audio. These accommodations were available only for students whose names were indicated when the lists of eligible students were submitted because of the additional time required to prepare them.

Other accommodations that were available to all students included:

- additional time
- one pause or several pauses during which students remained under supervision (assessment time does not include pauses)

Under no circumstances could school coordinators help students interpret the materials provided or guide their responses. Coordinators had to provide a description of any changes or irregularities to the test administration guidelines in the School Coordinator's Report.

## Questionnaires for the principal and teachers

Questionnaires for principals and teachers were delivered online and accessed with a unique login number and password. The School Questionnaire was usually completed by the school principal. The selected classes' mathematics teachers had to fill out the Teacher Questionnaire. In some provinces, there were a few schools that were structured in such a way that students were not registered or assigned to a particular grade. In this case, all mathematics teachers associated with the selected students were asked to complete a questionnaire (one questionnaire per teacher).

The questionnaires for the school principal and teachers were intended to establish links between the answers to the questionnaires and the students' performance. The data obtained from these also provided important information to those responsible for policy development. All questionnaire responses were confidential. The use of teachers' names was solely to link their ID number on a Student Questionnaire with that on a Teacher Questionnaire.

## Participation and exemption of students from the assessment

Grade 8/Secondary II students were expected to possess the necessary abilities to complete the assessment. It was therefore important that schools strongly encourage them to participate. While teachers could use various strategies to motivate the students to participate, they had to follow and comply with the assessment's administration procedures at test time.

It was possible, however, that some students would experience difficulty or great frustration participating in the assessment. For these students, teachers could predetermine that the assessment was not advisable in their case and exempt them. For example, students in the selected class with very limited mathematics, reading, or science skills could be exempted by the school from participating in the assessment. In some cases, the assessment might trigger emotional or physical reactions that staff in the principal's office considered harmful to a student. Regardless of whether a student participated in the assessment or was exempted for various reasons, the school coordinators had to indicate this using the participation codes in the *Handbook for Schools* and write the relevant code on the Student Tracking Form. It was important to assign a participation code to all selected students to ensure fair sampling for each province. There were eleven participation codes:

1 = Absent

2 = Participated during scheduled session

2A = Participated during scheduled session with an accommodation

3 = Participated during makeup session

4 = Exempted by the school

5 = Exempted because appropriate modifications could not be made

6 = No longer enrolled in this school/class

7 = Parents and/or students who do not wish to write

8 = Student not in Grade 8/Secondary II

9 = Home-schooled student

10 = Answer sheet and booklet were not returned; only the questionnaire data were available

11 = Student responded to fewer than 3 achievement items per domain and did not complete at least the first section (Section 1) of the contextual questionnaire

There were also three exemption codes:

F = exempted because of functional disabilities. A student who has a physical disability and who is unable to perform in the PCAP testing situation, even with one of the permitted accommodations should be exempted. A student who has a functional disability but is, nevertheless, able to participate should be included in the testing. The seven permitted accommodations were:

- additional time: although all students are allowed up to 30 additional minutes to complete the assessment, further additional time may be provided if the students receive such accommodations in a test situation during their regular school program
- a break, or multiple breaks, as long as students are supervised during the breaks
- an alternative setting
- use of Braille, large-print, coloured paper
- use of a scribe (writing verbatim: the scribe must write what student says without editing)
- verbatim reading of instructions only, for all domains
- verbatim reading of occasional prompts and/or questions for mathematics and science only (in cases where the entire mathematics and/or science portions of the test must be read, an audio version (on CD) can be provided)

I = exempted because of intellectual disabilities or socio-emotional conditions. A student who, in the professional opinion of the school principal or other qualified staff members, is considered to have an intellectual disability, or a socio-emotional condition, or has been tested as such, should be exempted. This category includes students who are emotionally or mentally unable to follow even the general instructions for the test.

N = exempted because of language (non-native speakers). This exemption is applicable only to those who do not have French or English as a first language. In large-scale assessments, schools can consider students who have been in Canada for less than two years as exempt.

The number and percentages of exempted students are indicated in Table A.2 in Appendix A of the PCAP 2019 public report (O'Grady, Houme, et al., 2021, p. 205).

## Organizing a makeup session

School coordinators had to ensure that the participation rate for students in their school was adequate. To this end, they had to count the number of 1 (absent) and 2 + 2A (participated during scheduled session) codes and calculate the percentage rate for student participation using the following formula:

$$\frac{(2+2A)}{(1+2+2A)}\times 100$$

If the student participation rate was less than 85 percent, a makeup session had to be held. The school coordinators were encouraged to include as many of the students who were absent as possible. If a student completed the assessment during the makeup session, their participation code changed from 1 (absent) to 3 (participated during makeup session) on the Student Tracking Form.

## Returning assessment materials

After assessing the students, the school coordinators filled out the School Coordinator's Report. They also filled out the School Packing List and indicated the number of each type of document being returned. As soon as possible following the assessment, they had to return to CMEC the School Packing List, the School Coordinator's Report, the completed Student Tracking Form, the School Questionnaires, the Teacher Questionnaires, the assessment booklets and answer sheets, as well as the copies and photocopies of unused assessment booklets.

## Coding session

The coding session for the main administration was held in Ottawa, for two weeks in July 2019. All the mathematics, reading, and science items were coded by teachers in the relevant domains. In all, there were 102 coders, both anglophone and francophone, in addition to two coding leaders for each domain.

In total, approximately 33,000 assessment booklets were scored using the online Central Coding System (CCS), with approximately 25,000 in English and 8,000 in French.

## Coding

All constructed-response items were coded by coders who were educators, because the questions required a degree of personal judgment and drew on their knowledge of the subject matter. Based on the descriptions in the coding guides, the coders assigned various codes to the students' responses and recorded them in the CCS. Once the coding session was finished, a database was created that contained all the assessment and questionnaire data.

## Coding Manual

In advance of the scoring session, coders were provided with a Coder's Manual that included information about the coding session's logistics and outlined the responsibilities of CMEC staff, coding leaders, and coders. It also provided information about how to handle special cases such as coder bias and suspected cheating. The Coder Feedback Form, which was to be completed at the end of the coding session, was also included in this manual.

## Coding Guide

The Coding Guide provided a general introduction to coding and detailed the principles of coding, such as guidelines for spelling and grammar errors and definitions of terms and special codes. The coding guide provided the classification for each question and a description of all possible codes as well as a range of sample answers that could be given full credit or partial credit for each question.

## Coding leaders

The coding leaders met for a few days in June to prepare for the coding session. They reviewed and adapted the materials related to the assessment, such as the Coding Guide. They also prepared the training materials for the coders. While preparing the training materials, coding leaders selected samples of student work to be used as examples or in training papers. Some samples selected during the field test process were also included in the training materials. The samples were used to show the distinction between the various codes for each item. Coding leaders were responsible for training table leaders and ensuring the smooth progress of the coding session.

## Table leaders

Table leaders led a table of six to eight coders. They were trained by the coding leaders. Their role included training the coders at their table, supervising their work, retraining individuals or groups as required to maintain coding consistency, and coding papers.

## Coder training

All coders, including table leaders, received training on the coding guides for mathematics, reading, or science, depending on their assigned coder role, before coding student responses. Prior to the training session, coding leaders selected student samples to be used in training. Examples were chosen to clearly illustrate the differences between the assigned codes for each question, and were reviewed and discussed. Training packages were then used to practise coding and to further internalize the coding scheme. Initially, pairs of coders worked collaboratively to code student responses until their coding was consistent with the coding guides. At the end of training, when coders were able to consistently apply the coding standards, they proceeded with individual coding until all student responses were coded. Coders were trained on one item and completed coding all student responses for that item before being trained on the next item. Throughout the coding processes, table leaders did a random check of the codes assigned by each coder to ensure consistent adherence to the coding guides. Issues that arose with respect to specific questions were addressed by either individual or group retraining or, in a few cases, by recoding the question.

Tables were assigned either English or French student responses to code. Tables of bilingual coders, who could help either the anglophone or francophone team with coding items, were assigned according to either English or French student responses, depending upon which team had more responses or was coding more slowly.

## Coding reliability

The goal of the reliability process was to provide evidence of the degree of agreement between coders for constructed-response items to demonstrate the consistent application of the coding guides. During the coding session, data were collected from reliability reviews and for inter-rater reliability or multiple coding.

### *Reliability reviews*

In a coding session, it is always important to implement the necessary procedures to ensure that coders are coding correctly, because they must all agree on the various codes to ensure the results' validity. Prior to the coding session, CMEC staff compiled several sets of responses at random for each item to conduct reliability reviews. The item set was then distributed to the coding leaders for coding. Their responses were returned to CMEC staff for entry into the CCS to be used for comparison with coder responses during the coding session. If coding leaders identified a specific issue arising with particular questions, additional reliability reviews were developed to target the issue. Reliability reviews thus functioned both as quality control and additional training for coders. Reliability reviews were run for all anglophone or francophone coders in all three domains. The reviews' goal was to monitor consistency throughout the coding session. The reliability reviews occurred approximately twice per day and followed this procedure:

- At a time determined by the coding leader, everyone stopped coding and coded the same set of student responses using the CCS.
- Codes from coders were compared to the benchmark (provided by the coding leaders).
- Results were immediately available to the coding leader.
- Coding leaders debriefed the entire group or individual coders.
- If the consistency was below 80 percent on a specific question, individuals or groups of coders were retrained and the student responses for that item were rescored as required.

The reliability reviews therefore checked the consistency between the experts' results and those of the coders. In other words, they checked whether the coders were assigning the same codes as the experts had for the items. For each reliability review, there was a percentage agreement calculated for each coder and each item. The level of agreement between the experts (coding leaders) and the coders was expected to be about 85 percent. If the overall reliability review was low for specific questions, then the group was retrained and previously scored material was rechecked by the coding leaders or the coding of the question began again. If the reliability review was low for specific coders or tables, then table leaders retrained the individual or the group of coders before proceeding with coding. Previously coded items by these coders were verified.

The percentage agreement by coder was determined using the following calculation:

$$Percentage\ agreement = \frac{Total\ number\ of\ agreed\ responses}{Total\ number\ of\ reliability\ tests} \times 100$$

At the end of the coding session, all the percentages obtained for each reliability review for each coder were compiled. This constituted the total level of agreement as a percentage. Results showed that most coders obtained a more-than-acceptable level of agreement with the experts. The reliability review results were satisfactory for all groups.

### *Inter-rater reliability (multiple coding)*

Multiple coding was a quality control measure in coding student responses for mathematics, reading, and science. For each item, 100 randomly selected responses were assigned to each coder. Coding leaders monitored the coder agreement and were able to identify issues throughout the coding process for each item and then retrain coders, either individually or in groups, as required.

### *Trend reliability*

Trend reliability was a quality control measure to estimate the degree of agreement between mathematics, reading, and science coders for the anchor items in PCAP 2019, PCAP 2016, and PCAP 2013. Four items for reading, eight items for mathematics, and six items for science were common among the three administrations. For each of these items, 385 student responses from PCAP 2016 were re-coded in 2019. Student responses were taken from Booklets 1, 2, and 4. Booklet 3 was not needed, due to the overlap in items across booklets. Student responses were scanned into the CCS. In this way, it was not possible for coders to identify if they were coding a response from 2016 or 2019, as 2019 also included students who completed the paper-based test. Trend-reliability coding was done throughout the main coding session.

## Reports and feedback

A variety of reports provided evidence of the program's strengths and weakness, which could be used to improve future PCAP administrations. School coordinators reported on the administrative process. This information was summarized and included in the summary reports by the provincial coordinators or submitted directly to the CMEC Secretariat. Coders provided feedback during the coding session. The information collected in these reports is summarized in the following sections.

### *Provincial Coordinator's Report*

Following the assessment's administration, the provincial coordinators drafted a report on the test's details and the information provided by the school coordinators. The report's purpose was to summarize schools' feedback about the test's administration. The information gathered from the provincial reports was used to make any necessary changes to the administration process for future assessments. The Provincial Coordinator's Report included seven questions.

First, the provincial coordinators had to summarize the methods the schools used to encourage students to participate seriously in the assessment. In most cases, the provinces sent information about the PCAP assessment to the parents or guardians of selected students to encourage them to participate. The school coordinators also met with the selected students before the test to discuss the purpose and importance of the assessment, to ensure students would make their best effort. They also let the students know that the assessment was anonymous and that their results would not be factored into the grade on their report card. At the end of the assessment, some students were rewarded for their participation — several schools offered a free breakfast, snacks, cafeteria coupons, etc. Some schools also provided external motivators such as gifts, certificates, or special privileges, or they thanked students by organizing events.

The *Handbook for Schools* outlined the administration procedures for the PCAP tests. Unfortunately, not all test administrators were familiar with this in advance. For example, teachers were encouraged to give students short breaks as required throughout the assessment, but not all teachers seemed to know about this accommodation ahead of time. The instructions indicated that students were permitted to use a calculator, manipulatives, and a dictionary (which could be French-English), or a thesaurus. Unfortunately, again, it seems that these things may not have been made available to students in all schools.

Provincial coordinators also had to summarize the problems encountered by the schools during the assessment's administration. Generally, technical issues related to the administration of the test occurred in only a very small proportion of schools.

The provinces' comments indicate that the majority of schools complied with the administration procedures. In most provinces, a high percentage of schools indicated that the assessment was administered in an excellent or a satisfactory manner. The schools that were only fairly satisfied with the test's administration mentioned that it was generally because they had received the administration materials late. Some schools also expressed concerns about specificity and clarity of the information related to the administration of the test.

It appears that the attitude of students who participated in the assessment was generally positive. Those who had a fairly negative attitude either did not see the value of the test, or were disappointed about missing activities, such as a sports event, to participate in it.

The school coordinators were generally satisfied with the *Handbook for Schools*. They said the information and instructions were clear and precise and that these documents facilitated the administration process. A few offered suggestions to improve future PCAP administrations. Since they found the amount of material too voluminous and detailed, some teachers suggested summarizing critical points in one or two pages and in very direct language with a point-by-point layout. Many also pointed out that the document needs to specify whether the use of calculators is allowed or not, and to provide specific instructions regarding extra test booklets for paper-based tests.

The coordinators' comments in the report were quite positive, and it appears that the administration process for the assessment proceeded smoothly. The suggestions and comments have been taken into account to improve the process for future assessments.

# 7 Online Assessment

## Item rendering and review

To ensure that the paper-based and online assessments were as similar as possible, items were developed, translated, and reviewed in a paper-based format before being imported into the testing platform. Selected-response items were formatted so that students could use radio buttons to select multiple-choice or dichotomous answers (e.g., true or false, yes or no). Several of the items developed for previous PCAP administrations, as well as new items for this cycle, required some form of animation. This included actions such as dragging and dropping to order responses, creating a line graph, or plotting points. To answer questions that required showing work, a mathematical keyboard was built into the system. In addition, the system included a calculator, a highlighting tool, and two tabs that students could click on to be provided with references (i.e., mathematical formulas) and a glossary of terms. Item review activities were conducted using a secure online review system developed by the platform provider.

## Computer requirements and test delivery

The basic hardware requirement for delivering the test was the availability of a suitable computer for each student. The computers were arranged so that the test could be supervised by a single test administrator, and in such a manner that students could not easily observe each other's screens. The requirements for compatible electronic devices are given in Table 7.1.

**Table 7.1 Computer requirements for PCAP online**

| Electronic Device | Requirements |
|---|---|
| Microsoft | Windows 7, 8.1, or 10 |
| Macintosh (Mac) | OSX 10.10–10.13 |
| Chromebook | ChromeOS 62+<br>Managed environment required |
| iPad | iOS 9.3.5+<br>Minimum screen size: 7.9-inch diagonal |
| Samsung or other Androids | 6.0+<br>Minimum screen size: 9.7-inch diagonal |

To determine the suitability of a computer for delivering the computer-based assessment, the PCAP assessment browser — i.e., the AWIS Secure Access Browser (SAB) — was downloaded on all the electronic devices used for administering PCAP in each school. The SAB included the appropriate default settings for each operating system. When operating, the SAB restricted student access to the internet and to other programs for the duration of the assessment. Students could use their own electronic devices, if such usage aligned with their school-board/district policy. After installing the SAB, each device was tested using the PCAP online practice tool to ensure that the assessment functioned properly. Schools that did not have enough compatible devices borrowed laptops from

CMEC to administer the PCAP assessment. For schools that did not have adequate internet for administering the test, CMEC provided portable wifi devices that operated using cellular coverage. The PCAP Helpdesk was available to provide technical support to schools during device verification and throughout the administration period.

Tests were administered securely over the internet and were protected through SSL 128-bit to 256-bit encryption by Verisign/Symantec. All servers were protected through a reverse security protocol Verisign/Symantec Certificate, similar to the security used for internet banking. Students accessed the test though a link on the CMEC website. A login form was provided for each class, with unique login codes and passwords for each student. In the event of system failure or disruption, the system automatically resumed the test on individual devices at the same point where the student using that device was interrupted.

Accommodations available to students were similar for both the online and paper-based tests, and, in advance of the assessment, schools identified students who required alternative formats. Students completing the online test had the option to use a pre-recorded voice or approved text-reader software (non-visual desktop access (NVDA, free, open-source software), VoiceOver by Apple, or ChromeVox by Google). Text-reader software that required students' devices to access outside programs was not compatible with the SAB because internet access was restricted during the assessment.

After the test administration, schools completed a session report form to provide information on technical problems, interventions, and any accommodations used for students.

## Data capture and coding student responses

Student responses were saved to the servers each time the student continued to the next question. This included using the Next or Back buttons, and the Review of Instructions screens. For those questions that took longer to complete, the system auto-saved every minute to avoid data loss.

After the assessment administration, open-ended responses from the online assessment were transferred into the online Central Coding System (CCS). Open-ended responses from the paper-based booklets were scanned and input into the online system.

Automated coding was done for selected-response, numerical-response, and drag-and-drop questions. Student responses to open-response questions were available in the CCS for human coding. For each domain, coders entered the system using a unique login code and password. The coder would see the list of questions that needed to be coded for each scenario. Coding was done by clicking on the radio button for the number that corresponded to the description of the code in the Coding Guide. Within the response area, coders could also add comments and send student responses to the coding leader if they saw an at-risk student response or inappropriate language, or if they suspected cheating.

The coding assignment was done using the CCS software. The quality of coding was monitored in several ways. Student responses were randomly distributed among coders assigned to each set of items. Multiple coding was applied to a subset of responses for which all coders were assigned the same set of student responses, which were randomly distributed throughout their allocated assignments. Multiple-coding reports were available at the item and coder level. Discrepancies between multiple-coded items and responses the coders deemed difficult or inappropriate were sent

for review by the table leader or coding leader. The CCS system provided real-time information on coding progress by item and by coder.

Another quality assurance measure was the use of reliability tests; approximately two reliability tests were administered for each item. Exemplars for the reliability tests were selected by the coding leaders and were also administered using the CCS. Reliability reports were available at the item and coder level. The results of these reliability tests helped determine whether codes had been applied correctly. This process allowed the early identification of coding issues and was followed by discussion and retraining, if required.

During the coding session, the table leaders spot-checked the work of coders each day. Spot-checking involved a review of codes assigned to responses. A general guide was that at least 10 responses per item were spot-checked. If a coder was uncertain about the code to assign to a particular response, the response could be marked for review and it would be sent automatically to the table leader or coding leader for advice.

Once the data collection was complete, the data were transferred to CMEC, using a secure file transfer protocol (FTP).

# 8 Studying the Transition to Online Delivery

PCAP 2019 marked the beginning of the transition to online assessment from a paper-based assessment. A pilot study was conducted to inform the delivery method for the test while a mode study was conducted during both the field test and main study for PCAP 2019 to determine whether trend items, which are those that had been administered in previous PCAP assessments, functioned in the same way when delivered online and on paper.

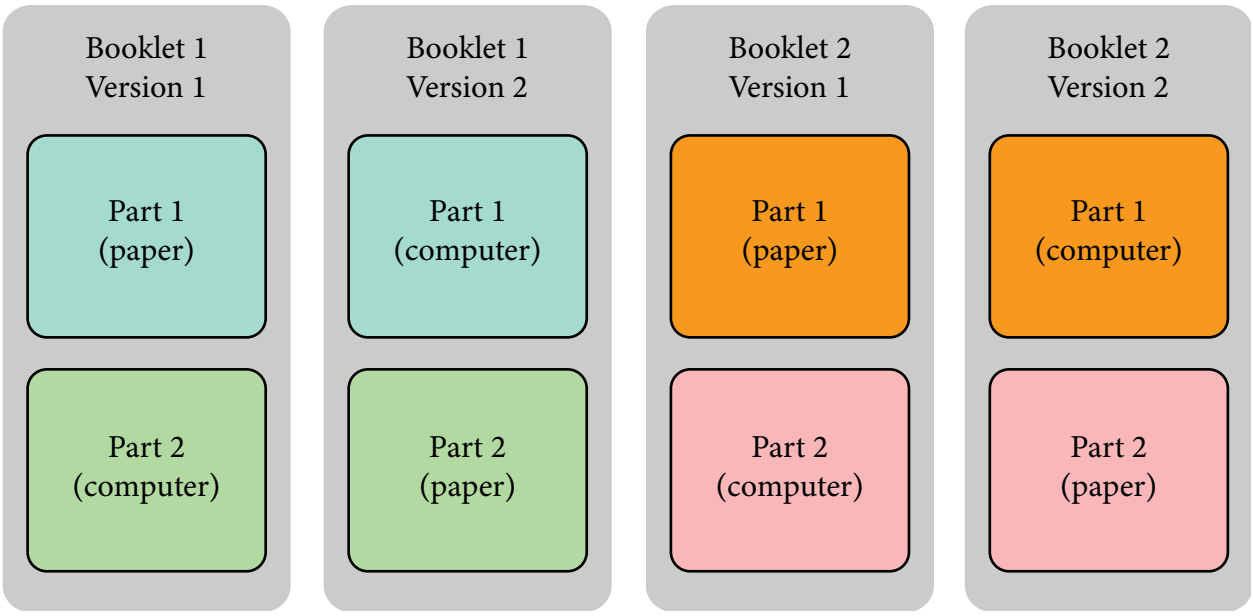
The pilot study was conducted in 43 schools in November 2016. Two vendors were selected for this study. One used a web browser–based test and the other a proprietary application (kiosk)–based test. The types of devices used for the test included PCs and Mac computers, Chromebooks, Android tablets, as well as other devices. The pilot study revealed that the delivery of the test using an online platform resulted in fewer technology challenges in schools. Based on the online pilot study results, it was anticipated that all schools would be able to meet the minimum technology requirements in order to administer the PCAP assessment online.

## Overview of the field trial mode study

The first mode study was conducted in spring 2017 using an online testing platform with a separate sample of 156 schools. A technology readiness survey was sent to these schools prior to the test administration, and a practice tool was available to the schools to ensure that equipment was configured properly.

The mode study booklets comprised items that were used as anchor or trend items in previous PCAP assessments in all three domains (mathematics, reading, and science). Two parallel booklets were designed with 33 to 34 items divided into two sections. The mode study was a 90-minute test conducted in two parts, with students writing a paper-based section followed by an online section, or vice versa, as shown in Figure 8.1.

**FIGURE 8.1 Booklet design for the field trial mode study**




Student results were aggregated at the test level to compare performance differences between students who wrote the same test items in different modes. For example, performance on Booklet 1, Part 1, Version 1 (paper) and Version 2 (computer) were compared to each other.

The mode study followed a one-group design with counterbalancing. Alternating the mode ensured that the order in which the paper mode and the online mode were administrated to the students did not affect the student scores. In that regard, the design facilitates a direct comparison between the two modes. For example, in a given booklet, Part 1 (paper) is directly comparable with Part 1 (computer), and Part 2 (paper) with Part 2 (computer). In each case, the groups satisfy the random group design.

Differential item functioning (DIF) analysis revealed significant difference between the two modes, especially in reading. In reading, higher mean scores were found in items administered online in Booklet 1, but the reverse was found for Booklet 2. In mathematics, average results were higher in both booklets administered on paper. Similar scores were found between the two test modes for science.

To further understand these differences, differential test functioning (DTF) was performed to determine if bias occur at the test level for each domain. Data for all booklets were combined and three DTFs were run (one per domain). Differential item functioning analysis system (DIFAS) software was used to perform the DTF. DIFAS examined the DIF effect variance as a measure of DTF, and DIF was estimated using the Mantel-Haenszel chi-square and Mantel's chi-square. The DTF results showed large DIF effect variances (large DTF bias) for mathematics and reading and small DIF effect variance (small DTF bias) for science.

Provincial curricula and assessment experts reviewed the results and identified items in the reading domain in which the layout of the reading prompt and the items were different between the two modes. Reading prompts occurred before the items in paper-based tests, so that student normally read the passage before seeing the items. In the online test, the format had items on the left of the screen and the passage on the right, so that students may have felt the need to search for specific information in the passage. Further, data were presented following the matrix sampling technique so that all items

within a domain were treated as a "single test." However, each "test" involved only small batches of items per student.

A second mode study was scheduled for the main study to further analyze the difference in item functioning between the two modes using a larger sample and addressing item layout issues where required.

## Overview of the mode study within the main study

During the PCAP 2019 main study, there were 29,926 participating students — 26,933 students wrote PCAP 2019 online, while a separate sample of 2,993 wrote the paper version of the assessment. The online and paper versions of the four booklets were as similar as possible. Differences were related to layout and tools. For example, in the online version, students used radio buttons to answer selected-response questions, had links to mathematics and science terms in English and French, and had access to calculators and graphing tools for specific questions. On the paper version, pull-out pages were used for selected-response answer sheets and for the translated mathematics and science terms. The Student Questionnaire followed the cognitive test in both versions of the test.

IRTPRO was used to perform the DIF analyses. This program uses the Lord's Wald test with accurate item parameter error variance-covariance matrices computed using the Supplemented EM (SEM) algorithm (Cai, 2008; Vector Psychometric Group, 2020.). Items were grouped by subject (mathematics, reading, and science). DIF analyses were performed for all items in each subject group. Difficulty parameters were compared first, by correlating the b-parameters between modes. Next, DIF results from the Wald test were examined, and for items showing significant DIF, the mode that was being "favoured" was noted. Student scores based on the IRT calibrations were also obtained to compare performance differences; for easier interpretation, the scores were scaled to the mean of 500 with standard deviation of 100.

Preliminary analysis showed that approximately half of the items showed DIF; however, the items were approximately equally distributed between favouring the online or the paper-based version.

# Setting a Performance Standard

Whenever tests' content and/or item types are modified significantly, standard setting is performed. If a given assessment does not change from one administration to the next, tests can be psychometrically equated (i.e., compared and adjusted statistically) so that students face the same performance standard each administration and are treated fairly. In 2019, mathematics was the major domain in PCAP for the second time. The assessment framework was updated and PCAP transitioned to online test delivery, so it was necessary to establish performance standards.

## Standard-setting sessions

Standard-setting sessions took place in February 2020 in Toronto. The meetings were divided into three sessions: a one-day leaders' training session; a two-day standard-setting session; and a one-day writing session to revise performance-level descriptors.

The standard setting aimed to articulate levels of performance on the PCAP mathematics assessment. These performance levels were delineated by cut scores that classified student performance. The standard-setting process was designed to produce these cut scores in a valid and systematic manner, first using a panel of content area experts and then including policy-makers and other stakeholders in the review phase. Two cut scores were set to differentiate between three levels of performance. Level 2 was designated as the acceptable level of performance for Grade 8/Secondary II students.

The standard-setting session was facilitated by Lennie Comeau from the Nova Scotia Ministry of Education and supported by two of his colleagues from the ministry and staff from the CMEC Secretariat. Participants took the tests, scored them, reviewed performance-level descriptors (PLDs), and then engaged in three rounds of test review using the bookmark standard-setting procedure (Cizek & Bunch, 2007). At the end of the three days, the cut scores recommended by the panellists were sent to the provincial coordinators for review. Procedures for developing and documenting those recommendations are spelled out here.

## Selection of an expert panel

It was important for CMEC that all provinces were involved and that they had an opportunity to participate in setting cut scores. Each province was invited to designate two representatives, one for each language sector, having some expertise in measurement and evaluation and in mathematics content for Grade 8/Secondary II. The standard-setting committee consisted of 27 panellists. CMEC solicited standard-setting panellists through nominations by the provincial coordinators. A key consideration of any such committee was that its members represent demographically relevant characteristics. To that end, CMEC constructed the committee to be appropriately balanced in variables like gender, experience, language, and geographical location. The 27 panellists also included teachers who worked with the target age group. CMEC took care to ensure robust representation of both English and French speakers.

## Preliminary performance-level descriptors

Important to any standard-setting process are performance-level descriptors, or PLDs, which describe what students should know and be able to do at each of the proficiency levels within Grade 8/ Secondary II. The PLDs are crucial to the standard-setting process because they provide guidance to panelists by helping them conceptualize differences in performance levels among students.

The PLDs from PCAP 2010, when mathematics was the major domain for the first time, were reviewed. These were statements describing what students at the four performance levels knew and could do and were referred to by panellists throughout the standard-setting process so that they had a solid working concept of what student performance should be at each proficiency level.

## Security of materials

Because standard setting uses operational materials, security was crucial. Upon signing in to the workshop, each panelist received a unique identification code. All secure material contained the same codes so that upon distribution the number on the material matched the panelist ID number. Panelists were informed that it was their responsibility to ensure that the material with their number remained confidential. Panelists were also asked to sign a nondisclosure form prior to receiving any secure material. No material was allowed to leave the breakout rooms at any point during the day.

## The bookmark procedure

The bookmark method was selected to maintain continuity with prior PCAP standard-setting sessions for the following reasons: the method can accommodate mixed-format assessments; it lets participants review selected-response and constructed-response items together; and it is based on, and ideally suited for, item response theory (IRT)–based assessment approaches. The bookmark method requires fewer and simpler decisions from participants than other standard-setting methods. For these reasons, the bookmark method was considered an efficient, effective, and appropriate approach for standard setting for PCAP.

The overall format of the PCAP 2019 assessment was a mix of selected response (e.g., multiple-choice (MC), true or false, and yes or no) with a significant number of short-constructed-response (SCR) items and extended-constructed-response (ECR) or open-ended (OE) items. SCR items were mathematics items that could be answered with a brief response that was scored dichotomously (coded 1 or 0 for correct or incorrect, respectively). ECR items were one-, two-, or three-point items that required a student's longer written response and for which partial credit could be given.

With the bookmark procedure, panellists examined test items in an ordered-item booklet (OIB) in which all the items from all four booklets used in the assessment were arranged in order of difficulty, with the easiest item placed on the first page and the most difficult item on the last page. MC and SCR items appeared only once in the OIB, but ECR items and context information appeared once for each score point. An item worth two points appeared twice, the first time with a sample response representing one point, then later with a sample response representing two points. An item worth three points appeared three times in the OIB. Each page contained essential information about the item, including its position in the OIB, its position in the original booklet or booklets, its classification by mathematics subdomain, the item difficulty, and the score point associated with

the item in that position. The 2PL IRT model and the Generalized Partial Credit (GPC) model are typically used by CMEC on item calibrations and test construction. The item difficulties were indicated by the b (location) parameters from the calibrations.

During the review of the OIB, panellists were asked to identify the item where students at each performance level would have a two-thirds chance of getting the item correct, and mark that item on the OIB with a bookmark — hence the term bookmark method. Items before the bookmark reflect the test content that students at the proficiency level should master, and items after the bookmark should reflect the test content that is difficult for the student. Each time the panellists reviewed an item, they were asked to think about the following question: "Would the student at this proficiency level have a two-thirds chance of answering this item correctly?" If the panellist answered yes, they moved on to the next item. If they answered no, they bookmarked the item. The RP67 was calculated to obtain the cut scores between proficiency Levels 1 and 2, between Levels 2 and 3, and between Levels 3 and 4.

## Standard-setting procedure

Twenty-seven participants from all provinces took part in the standard-setting session. Participants were assigned to three anglophone or two francophone tables. Each table had a leader and five or six participants. The cut-score-setting process took two days, with a third day set aside for refining performance-level descriptors.

The panel was given a presentation on PCAP, administration procedures, item characteristics, and the assessment framework (this information was especially relevant for participants who were taking part in a CMEC pan-Canadian assessment-related project for the first time), as well as cut-off points, the bookmark method, performance levels, the session schedule, keys for selected response items, and coding guides constructed response items. Most participants had never used the bookmark method and required briefing on the process and their tasks over the two days of the session. Finally, information was provided on performance levels to help the panel clearly distinguish among the four levels.

Participants then took the rest of the morning and part of the afternoon to become familiar with the assessment instrument and the materials for the session. This step took some time, but it was necessary for panel members to review the materials carefully to gain "fluency" with the assessment. Participants had discussions at their tables concerning the assessment items and item difficulty, and they were given an opportunity not only to review but also to answer items and score their answers, thereby gaining insight into performance-level descriptors.

The bookmark procedure involves three rounds, the first two of which are practise rounds to become familiar with the process. The first bookmarking round took place before the end of the first day, with participants reviewing each item in the OIB independently. Each participant selected a cut-off point or cut score — the last question that a student had a two-thirds chance of successfully answering for a given performance level — and placed a bookmark in the OIB. The groups then discussed their conclusions and the reasons that one item was more difficult than ones ranked lower in the booklet. Following discussion, all participants' responses were recorded in an Excel file, with the OIB page numbers denoting the three cut scores. The mean of all responses defined the location of cut scores between Levels 1 and 2, between Levels 2 and 3, and between Levels 3 and 4.

The second day began with a full group discussion on the first bookmarking round. The results were posted with a graph that illustrated all the OIB page numbers identified by the panellists at the three cut scores, and a table of statistics at each cut score with the mean difficulty, the mean page numbers, and the range of page numbers. Major variations were evident between participants' responses. These variations led to important and relevant discussions, with panel members explaining to each other why they had bookmarked a specific item.

The second round was similar to the first, with panellists placing bookmarks independently in the OIB to determine the three cut scores and providing a rationale for their choices as a group. The goal of the second round was to obtain a more coherent set of results than had been gathered in the first round. The results were compiled and shared with the group. Some participants decided to change their bookmarks, while others chose to leave them on the same item. There were fewer variations in responses for the cut scores than in the first round. For the second round, participants were shown impact data, that is, the percentage of students performing at each performance level. Showing participants the impact data allowed them to check their choices against the outcomes and readjust their cut-off points accordingly. This is important because the panellists were educational experts with working knowledge about the expected distributions of students' performance. IRT cumulative frequency tables for the theta statistic for each booklet were compiled ahead of time and used during the sessions to determine the proportion of students who would fall below and within each of the cut-level groupings.[7] However, the panel was clearly instructed to place bookmarks based on item difficulty and not on the percentage of students that participants wished to assign to each level.

In the third round, participants bookmarked the cut score between the levels for the last time in the OIB, either maintaining or changing their previous choices. Results of this round were much more coherent than results from the previous two rounds. Based on panel responses in the third round, the percentages of students at each performance level were determined as shown in Table 9.1.

**TABLE 9.1 Distribution of students by performance level in mathematics**

| Performance Level | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|
| **Mean score in mathematics** | 385 or below | 386 to 497 | 498 to 644 | 645 or above |
| **Percentage of students** | 11 | 39 | 42 | 8 |

A questionnaire was distributed to participants at the end of the session to collect information, comments, and feedback on the standard-setting process and the method used, as well as on the assessment instrument itself. Most panel members reported that they had enjoyed the session, that they had been comfortable with the process, that the session had been an enriching experience, and that the bookmark method was a fair and easy-to-understand way to set cut scores. The majority of participants also appeared satisfied with the organization of the session and with the leaders and facilitators, and they commented favourably on the assessment instrument. Most stated that the scenarios and questions were appropriate for and fair to Grade 8/Secondary II students.

[7] The theta statistic was adjusted for the two-thirds response probability as described earlier.

## Performance-level descriptors

Following the standard-setting process, a subset of panelists revised the performance-level descriptors. They examined all items within the range of scores that defined the four levels of performance. Using these items, they revised the description of the knowledge and skills that characterized achievement at each of the four performance levels.[8] Based on pan-Canadian curriculum expectations in mathematics, the expected level of performance of Grade 8/Secondary II students is Level 2. Students achieving at Level 1 are below what's expected of students in their grade.

Performance levels are thus summarized as the percentage of students reaching each level. Tasks at the lower end of the scale (Level 1) are deemed easier and less complex than tasks at the higher end (Level 4), and this progression in task difficulty/complexity applies both to overall mathematics and to each subdomain in the assessment.

[8] These descriptions appear in the PCAP 2019 public report (O'Grady, Houme, et al., 2021).

# 10 Processing PCAP Data

Data processing is an important and fairly complex part of the project, because specific steps must be followed to ensure valid results. CMEC convened a Technical Advisory Committee — a group of experts in measurement and assessment, as well as in statistics — recognized in their respective fields throughout Canada, with broad expertise in large-scale education assessments. The goal was to seek their input and any advice that could help enhance the analyses that were conducted.

## Data cleaning

When data were received after the coding session, the first step was to check the consistency of the database structure with the CMEC database. The data officer identified all the variables, adding or deleting variables as necessary. Consistency checks were completed for participation codes, achievement data, questionnaire data, and the data received from the data entry company. All deviations were checked and verified. The data files were then used for specific data cleaning and recoding procedures.

### *General recoding*

After the CMEC data centre had investigated all deviations and introduced corrections into the database, the following general rules were applied to the unresolved inconsistencies in the PCAP database (this was usually a very small number of cases and/or variables per province, if any):

- Unresolved inconsistencies regarding student and school identification led to the deletion of the record in the database.
- Student records that did not contain both achievement and questionnaire data were corrected with the appropriate participation code. This also applied to students who responded to fewer than three achievement items per domain and did not complete at least the background part (Section 1) of the contextual questionnaire.
- Duplicate data records were identified and only one record was kept, if the two records showed a 100 percent match of identical data. For duplicate records that did not match, efforts were made to refer to the booklets for clarification and correction.
- On very rare occasions where the source of inconsistencies could not be identified, both duplicated records were deleted.

### *Review of the sampling data*

The final data-cleaning step in sampling and tracking data was based on the analysis of tracking files (e.g., Student Tracking Form, Booklet Tracking Form). CMEC analyzed the sampling and tracking data, checked them, and, if required, completed further recoding. For example, if a province had greater numbers of students in one language than required by the sampling framework, then the language codes for schools were verified and recoded as necessary.

## Final review of the data and preparing the database

Once all the data were captured and reviewed, the files were compiled and merged. The finalized databases were then used for preliminary analysis and weighting. For the questionnaires, the reports contained descriptive statistics on every item in the questionnaire. For achievement data, classical analysis and differential item functioning (DIF) analysis were conducted. These analyses provided information about test items that appeared to have behaved differently and about any ambiguous data remaining in the questionnaires. With such information, the key was corrected, if necessary, and ambiguous data were further recoded. For example, if an ambiguity was a result of an error in printing, layout, or translation, then a "not applicable" code was applied to the item.

Recoding (required as a result of the initial analysis of achievement and questionnaire data) was introduced into the data files. Student, teacher, and school weights were estimated by Statistics Canada simultaneously based on the sample size allocations and non-response adjustments. Appendix A provides more detailed information on weighting. Upon weighting by Statistics Canada, weights were sent to CMEC, and the final weights were used for further analyses and linking of the assessments.

# 11 Analysis of Achievement Data

This chapter outlines the PCAP 2019 analysis of achievement data. It identifies, describes, and gives a detailed schedule for how the tasks were performed and coordinated. The analysis plan included the following:

1. preliminary analysis
2. item analysis
   i. classical analysis
   ii. IRT analysis
   iii. differential item function (DIF) analysis
3. test functioning
4. linking and equating PCAP 2019 mathematics, reading, and science with PCAP 2007, 2010, 2013, and 2016
5. coding and scaling PCAP 2019 performance data
6. standard error estimates
7. presenting the PCAP 2019 performance results

## Preliminary analysis

The preliminary analysis was an extension of the data-cleaning process. It included three steps: (1) data screening, (2) item recoding, and (3) missing-data handling. These steps were performed for each booklet with breakdowns by province and by language. These breakdowns facilitated the data-checking process, for example, identifying cases of interest regarding items that a student did not reach.

### *Data screening*

Frequency tables were produced for each item with breakdowns by province. They were used:

- to check for anomalous data (e.g., outliers, incorrect keys, etc.);
- to examine (first-level examination) the distribution of the responses; and
- to determine (and eventually assess) the missing rate per item and per booklet.[9]

In addition, data were cleaned so that cases with all blank responses were removed. During this stage of cleaning, cases were identified where students who wrote the assessment were exempted (assigned participation codes 4, 5, or 6 — that is, the student was exempted by the school, exempted because appropriate modifications could not be made, or no longer attended the school). On the other hand, some of these students also wrote with accommodations. Further investigation revealed that some of these students had enough data to be retained, meaning they had attempted at least

[9] Missing data types and treatment are described later.

one item per question type per subject. Cases of misidentification could have been the result of the student being given a paper-based booklet with an ID number that did not correspond to the Student Tracking Form or a student being given a second login if a technical problem occurred with the online assessment. Upon consulting with the provinces, it was decided to keep these students in the final data set. However, data for these students were not included in the calibration process. Verification of participation rates for all provinces is done during data screening.

## *Item recoding*

Prior to the data analyses, the PCAP 2019 raw data sets were recoded and cleaned based on the analyses required. Data recoding was required for the identification of valid items, recoding of different types of missing responses, and IRT analyses. The PCAP 2019 raw assessment data included both valid and invalid responses to the test items. For a multiple-choice (MC) item in the paper-based assessment, a response was valid if the student chose only one response option, whether the choice was correct or not. The answer was considered invalid if more than one option was selected. This issue was not a consideration for the online assessment. The student's constructed response (CR) to an open-ended item was treated as valid if it was related to the question being asked, regardless of whether it deserved no credit or partial or full credit. If the student's response was unrelated to the question, it was considered incorrect. Numerical codes were assigned to invalid MC and CR items.

The MC items in the English and French versions were separately recoded. This was necessary because the keys for some of the MC reading items, which were anchors from previous assessments, were not the same in both languages because the distractors appeared hierarchically in the form of a pyramid. Failure to recode these items could have led to problems during the calibration process (e.g., convergence would not be achieved).

Each response option was transformed into a variable with binary values. Four new dichotomous variables were derived for each MC item. The new set of variables included one variable for the correct response and one variable for each of the three distractors. These variables were used for the classical item analysis, especially to assess the behaviour of the distractors.

## *Missing data*

As is the case in other large-scale assessments, four types of data were missing from the PCAP 2019 assessment:[10]

- missing due to item sampling (not administered);
- missing because a student did not see the item (not applicable);
- missing response because a student runs out of time to complete the assessment (not reached); and
- omitted items (omitted).

To distinguish these types of missing data from each other, and from multiple responses or invalid responses, the following codes were used:

[10] PISA added multiple or invalid responses as a fourth category of missing data. Multiple responses were not considered as missing data in PCAP and were treated as different types of data.

- not administered: system missing
- not applicable: 7
- not reached: 6
- omitted: 9

### Not-administered items

Not-administered items stem from the PCAP assessment design that relies on the multiple-matrix sampling technique. This technique divides the assessment items into sections or booklets with some items that are common to some or all of the sections. Each section is then assigned to a distinct sub-group of the main sample. In PCAP, the questions were divided into four booklets, with some clusters of items that were common between pairs of booklets. Since each student was administered only some of the test items, there were no responses for items assigned to the other three booklets and so responses were missing because of the assessment's design. Therefore, not-administered items fell into the category of data that were missing completely at random (MCAR). As such, they can be ignored and were treated as missing data.

### Not-applicable items

The "not-applicable" code was used if a question was misprinted or displayed improperly on a digital device, making it impossible for the student to answer. For example, there may have been a printing or screen resolution error so that the question was not legible. The not-applicable code was used in only a few cases and these were treated as missing values.

### Not-reached items

Not-reached items correspond to non-answered questions that were clustered toward the end of an assessment. They occur in a student's vector of responses because the student didn't have time to provide an answer to them. In international assessments, an item is considered not reached when the item itself and the one that immediately precedes it were not answered or if the examinee attempted no subsequent items in the remainder of the booklet.[11] In other words, the first item with a missing response following the last valid (or invalid) answer was treated as the one the student was attempting but didn't have time to complete.

Not-reached items in PCAP were treated as ignored. This method is supported by Lord (1980), who argues that readily quantifiable information from such items can't be obtained for person location (see also de Ayala, 2009). PCAP 2019 treated not-reached items following the approaches used by TIMSS and PIRLS. These two international assessments treat them as not-administered when calibrating items. However, when estimating theta scores, these items are treated as incorrect responses.

### Omitted items

The omitted items were skipped throughout the assessment either inadvertently or because the student didn't know the answer. These items appeared earlier in the test as opposed to not-reached items that were clustered toward the end. Lord suggests that omitted items should not be ignored

[11] Not-reached items are defined in the PISA, PIRLS, and TIMSS technical reports. In PIRLS and TIMSS, "an item is considered not-reached when the item itself and the item immediately preceding it are not answered, and there are no other items completed in the remainder of the booklet" (Foy, Brossman, & Galia, 2012, p. 18). In PISA, not-reached items are "all consecutive missing values clustered at the end of a test session except for the first value of the missing series, which is coded as missing" (OECD, 2012, p. 199).

(cited in de Ayala, 2009, p. 150). He argues that with the practice of ignoring omitted items, a higher proficiency estimate could be obtained if a student responds only to questions they have confidence in correctly answering. Even though PCAP does not report individual scores, omitted items received the same code as an incorrect response.

### Invalid response

Invalid responses occur when the respondent chooses more than one answer for a given item in the paper version of the assessment. These types of response were coded 8.

## Item analysis

Two families of analysis were run: (1) classical test theory item analysis and (2) IRT analysis.

### *Classical test theory item analysis*

The objective of the classical analysis was to produce statistics for a second review of the PCAP 2019 items. For the major domain, which was mathematics, the first review used the field test data. The field test for mathematics consisted of new items. For the main study, mathematics consisted of both new items and anchor items, while the minor domains, reading and science, consisted of only anchor items. Anchor items were from previous administrations when the domain was the focus of the assessment (reading in 2007 and 2016, mathematics in 2010, and science in 2013), and were used to assess the change in item performance over time (or from one cohort to another) on the basis of their estimated difficulty. The statistics were reviewed in preparation for the selection of items to be included in PCAP 2019.

The classical test theory item analysis for the major domain items focused on the following:

- item difficulty
- item discrimination
- specific statistics for the selected-response (SR) items (e.g., multiple choice, true and false, yes and no)
- specific statistics for the constructed-response (CR) items
- percentage of students choosing each response option for each item
- percentage of students not reaching the item
- percentage of students omitting the item
- reliability indices (i.e., the internal consistency index for the SR items and the intercoder agreement for CR items)

These statistics were computed for each booklet — four booklets for the English version of the test and four booklets for the French version — and for each of these booklets in both the online and paper versions of the test. For the minor domains, the placement of the items in PCAP 2019 was consistent with their position in the original assessment. Nonetheless, their position's effect was also assessed.

### Item difficulty

For each SR item and for dichotomous CR items, the difficulty corresponded to the classical *p*-value. For polytomous CR items, the average percentage reflected their difficulty. In both cases, not-reached responses were excluded from the calculation.

### Item discrimination

For both SR and CR items, the corrected item-total correlation — that is, the relation between the correct response to an item and the total score — was computed. A moderately positive correlation between items with good measurement properties and the scale was expected and achieved irrespective of the mode of administration. Not-reached responses were excluded from the calculation.

### Specific statistics for MC items

For multiple-choice items, the specific statistics included:

- the percentage of students choosing each distractor
- the point-biserial correlation between each distractor and the total score on all the items administered to a student for a given domain. For items with good measurement properties, distractors exhibited negative correlations.

### Specific statistics for CR items

For items that required constructed responses, the specific statistics included:

- the percentage of students responding at each score level
- the point-biserial correlation between each score level and the total score on all the items administered to a student, for a given domain. This correlation was expected to be increasingly ordered from negative to positive by increasing score increments for items with good measurement properties.

### Examining for missing data

The following were examined for each item:

- percentage of students omitting the item
- percentage of students not reaching the item
- point-biserial correlation between the omitted variable of the item and the total score on all the items administered to a student for a given domain
- the point-biserial correlation between any not-reached variable of the item and the total score on all the items administered to a student for a given domain

All these statistics were also estimated for each population (or province, if only one language group was reported) for comparison with the pan-Canadian-level estimates.

### Reliability of the PCAP 2019 assessment

For each domain and subdomain, Cronbach's alpha was used as the internal consistency index. It was computed across all assessment booklets as an index of reliability. The means of this reliability index for each domain and subdomain were also computed. The same was done for each province.

### Problematic items

Problematic items were flagged based on the classical analysis. An item was flagged as problematic if one or more of the following conditions were present:

- point-biserial correlation less than 0.20
- p-value less than 0.20
- p-value equal to or greater than 0.85
- items easier or more difficult for a province relative to the national average[12]
- positive point-biserial correlation for more than one distractor in an MC item, or point-biserial correlations across levels of constructed response items not ordered
- less than 5 percent of students selecting one of the MC distractors
- less than 10 percent of students being awarded the score value for a CR item
- intercoder agreement of less than 70 percent on the score value of a CR item

## IRT analysis

The IRT analysis process involved: (1) assessing the dimensionality, (2) estimating items' parameters, and (3) assessing the IRT model fit. Analyzing the fit of IRT models included the local item dependence (LID), the agreement between the model's mathematical function and the data, and the invariance. The process ended with assessing the differential item functioning (DIF) of the items as part of the validity evidence.

### *Assessing the dimensionality of PCAP 2019*

The dimensionality was assessed by item factor analysis (IFA). The IFA designates the class of nonlinear approaches to determining the factorial structure of categorical data (Cai, 2010). These approaches are more appropriate than the classical factor analysis, which is based on a matrix of linear correlation between the observed variables. As a linear approach, it leads to extracting possible artifactual factors when dealing with dichotomous (or polytomous) variables (de Ayala, 2009; Laveault & Grégoire, 2002). Nonlinear approaches are, therefore, more in alignment with these types of data than the linear approaches (McDonald, 1967).

The statistics program IRTPRO implemented a full information maximum likelihood (FIML) procedure that accounted for the nonlinearity between the observed variables and between the observed variables and the construct under consideration. It was concluded that the unidimensionality assumption for the major domain was satisfied.

[12] This assumes that the Rasch model is fitted to the data as a means for flagging items and that the item by province interaction analysis is run.

## Item calibrations and assessing the fit of IRT models

Items from pairs of booklets were calibrated concurrently to link all the booklets and to put scores on a common metric. This procedure makes it possible to estimate theta scores in a way that does not depend on the set of items to which the students responded. Items from the three domains were calibrated independently as they were measuring different subjects.

Three IRT models were fitted to the data simultaneously. For the MC items, the modelling fit the two-parameter logistic model (2PLM) to the data. It was then compared to the three-parameter logistic model (3PLM). The 2PLM was retained because the model fit didn't improve significantly when 3PLM was tested; for the dichotomous CR items, the 2PLM was used. The polytomous CR items were calibrated using the Generalized Partial Credit Model (GPCM). For the estimation of all three item parameters, the Maximum Marginal Likelihood (MMLE) method was used. The model fit assessment involved assessing the local item dependency (LID), the agreement between the distribution of the empirical data, and the theoretical (or expected) distribution.

The LID was assessed by means of LD $\chi^2$ statistic (Chen & Thissen, 1997). This statistic is computed by comparing the observed and expected frequencies in each of the two-way cross tabulations between responses to each item and each of the other items. These diagnostic statistics are (approximately) standardized $\chi^2$ values that become large if a pair of items indicates local dependence, that is, if data for that item pair indicate a violation of the local independence.

The adequacy of the specified mathematical function to the actual data shape was assessed based on the S-$\chi^2$ statistics (IRTPRO does not produce and does not endorse producing the empirical item response curve). The S-$\chi^2$ statistics are based on the difference between observed and expected frequencies in response categories by summed scores.

## Invariance of PCAP 2019

As part of the mode study, differential item functioning (DIF) for the paper mode and the online mode of administering PCAP 2019 was investigated. More specifically, the analysis sought to determine to what extent the results from the new format, the electronic mode, were comparable to the paper mode, and thus to previous administrations. To that end, it was important to investigate whether the PCAP 2019 items show similar psychometric properties across modes. Of particular interest were the trend items, as it was crucial that they operated equivalently across modes in order to maintain the assessment of change over time.

This investigation was done through a set of DIF analysis: two IRT-based DIF procedures and a non-IRT-based technique. The first analysis was the Wald test–based approach implemented in IRTPRO (IRTPRO Wald test) for DIF assessment. Cai (2010) pointed out that these tests are modelled on the Lord (1977) proposal, "with accurate item parameter error variance-covariance matrices computed using a supplemented expected maximum algorithm" (Vector Psychometric Group, 2020). Given the large size of the PCAP assessment, it was expected that the approach would result in a sizable number of items with statistically significant DIF. However, the outcome of the first-round analysis was used for further investigation. Therefore, two additional approaches were used for the subsequent assessments: logistic regression analysis of DIF with effect size (LRDIF_ES) computation and a logistic ordinal regression differential item functioning (lordif) detecting based on the IRT framework (Choi, Gibbons, & Crane, 2016a). The LRDIF_ES was performed using a SAS macro that can simultaneously handle dichotomous and polytomous items (Fu & Monfils,

2016). The lordif detecting used a R-package of the same name that also can simultaneously manage binary and polytomous items (Choi, Gibbons, & Crane, 2016b). The two approaches have the advantage of computing the effect size, a more useful statistic to appraise the magnitude of the difference in item functioning. They also allow the investigation of both uniform and non-uniform DIF through a single omnibus test. This involves fitting and comparing, for each item, three models presented in a hierarchical and ascending order. These are, from the most parsimonious to the most complex, the no-DIF, the uniform DIF, and the non-uniform DIF models. They are hereafter referred to, respectively, as Model 1, Model 2, and Model 3. Model 1 posits that no DIF is present; only the latent trait predicts the probability of a student's score on a given item. Model 2 tests a main effect term; it postulates that, in addition to the latent trait, group membership independently predicts the probability of a student's score on a given item — that is, a DIF is present and is consistent on the continuum of the trait levels. Model 3 adds an interaction between the two predictors — that is, the group membership effect varies, conditional on the trait level.

In addition to these common features and capabilities, the LRDIF_ES enables the researcher to include external explanatory variables beyond the ability and the group membership in its modelling (Fu & Monfils, 2016). For PCAP 2019, background variables such as parent education level and numbers of books at home, reflecting a measure of socioeconomic status, were included in the logistic model as external variables. The LRDIF_ES also includes "an option to produce a line plot for each item and grouping variable, where the mean item score and the lower and upper boundaries of the 95% confidence interval of mean score for each group are plotted against the criterion (matching) variable" (Fu & Monfils, 2016, p. 1). While the LRDIF_ES relies on the summed trait score, the lordif has a further advantage of integrating both the logistic regression and the IRT techniques in one framework. Thus, the item parameters and the latent trait are estimated through IRT techniques. To be consistent with the IRTPRO modelling, the generalized partial credit model was fitted to the PCAP 2019 polytomous items instead of the default graded rating modelling implemented in lordif. The lordif also implements the Stocking-Lord equating technique to enable the investigation of the DIF impact on the same metric. In addition, it utilizes a Monte Carlo simulation to derive empirical threshold values for various DIF statistics and magnitude measures.

As one can note, the approaches offer a good supplement for a more extensive way to evaluate the practical significance of the PCAP 2019 DIF items. In both procedures, a likelihood ratio test was used for model comparison. A significant difference between Model 1 and Model 2 indicated a presence of uniform DIF. If the likelihood ratio test resulted in a significant difference between Model 2 and Model 3, it signalled the presence of non-uniform DIF. An overall or total DIF was tenable if the result of the likelihood ratio test comparing the first and the last models was significant. While the likelihood test is useful and commonly used for model comparison, it suffers the same limits as most statistical tests, as it is sensitive to the sample size. For large sample sizes, it tends to yield statistically significant results that may be practically negligible. Therefore, the effect size statistics were used to assess the magnitude of the PCAP 2019 DIF items. To that end, the assessment relied on the NagelKerke $R^2$, McFadden $R^2$, and the regression coefficient ($\beta$). More specifically, the size of the change, in these indices ($\Delta R^2$ and $\Delta\beta$), from a more parsimonious model to a more complex one, revealed the magnitude of DIF. For both dichotomous and polytomous items, the magnitude, as related to $R^2$, was classified using the criteria recommended by Educational Testing Service (ETS) in the Mantel-Haenszel procedure (Fu & Monfils, 2016; see also Choi, Gibbons, & Crane, 2011; Jodoin & Gierl, 2001). Per ETS, a DIF item falls into one of the three categories shown in Table 11.1.

### Table 11.1 Education Testing Service (ETS) DIF classification

| Category | Criterion | Magnitude |
|---|---|---|
| A | $\Delta R^2 < .035$ | negligible DIF |
| B | $.035 \leq \Delta R^2 < .070$ | moderate DIF |
| C | $\Delta R^2 \geq .070$ | large DIF |

For Δβ, known as the proportionate change in point estimates for the beta (proportionate change in beta, for short), various criteria have been proposed to conclude a practically meaningful impact of a DIF item. For instance, Maldonado and Greenland (1993) suggested a threshold of 10 percent change, while others have suggested lower thresholds (e.g., 5 percent) for this effect size measure (Crane, van Belle, & Larson, 2004).

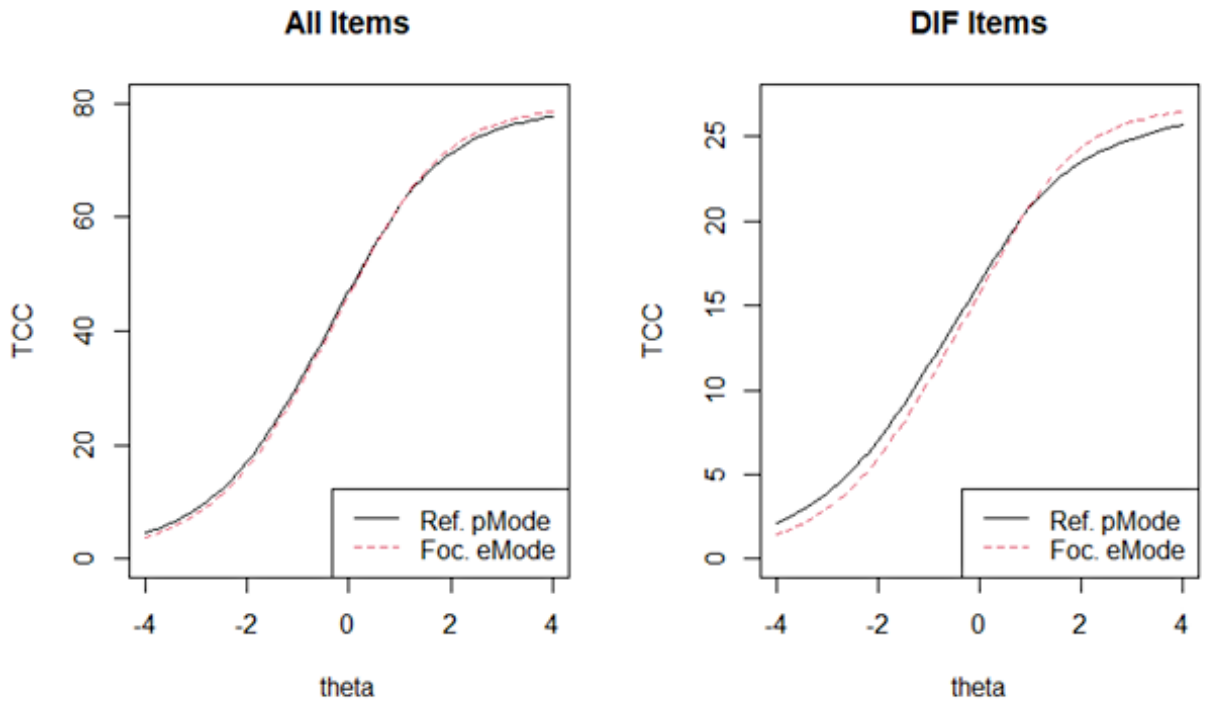
The results in Table 11.2 indicate a negligible-to-moderate magnitude of differential functioning of the PCAP 2019 mode DIF items. The effect size measures $\Delta R^2$ all fell below .05. In fact, the inspection of the results by items (not reported here) indicated that only one item exhibited moderate DIF ($\Delta R^2 = .036$). Similarly, all the values for the proportionate changes in β fell below .01, except for one item where the measure equalled .07. It should be pointed out that the effect size statistics aligned with the graphical depicture of the DIF magnitude, as evidenced by the test characteristic curve (TCC) illustrating the differential test functioning. For mathematics, there was almost an overlap of the online and the paper modes TCCs (see Figure 11.1). For reading and science, the TCCs were very close, and therefore were not concerning. It is important to note that all of the trend items, in all three domains, displayed marginal effect sizes. The findings were shared with the Technical Advisory Committee. Based on these findings, it was concluded that the detected DIFs were practically insignificant. In other words, using data from the online mode of administration would be inconsequential for trend analysis and comparison with past PCAP assessments.

### Table 11.2 Ranges of effect size measures for mode DIF items

| Domain | $\Delta R^2$ | Δβ |
|---|---|---|
| MathN* | .001–.036 | .001–.070 |
| MathT* | .001–.004 | .002–.003 |
| Reading | .001–.004 | .001–.003 |
| Science | .001–.009 | .001–.013 |

*MathN=Mathematics new items; MathT=Mathematics trend items

**FIGURE 11.1 Mode DIF detection: Test characteristic curves for mathematics items**




The analysis that follows deals exclusively with the online administration data, as these are the data to be reported moving forward. Before generating the final ability trait on which to report, another DIF investigation was conducted for language and gender. It involved assessing the extent to which some of the PCAP 2019 items displayed different statistical proprieties (e.g., level of difficulty) for language and gender. As in the assessment of invariance across mode, the Wald test DIF detection, the LRDIF_ES, and the lordif procedures, as described above, were applied to the data. While the results showed that some of the items exhibited DIF in language and gender, the effect sizes were not large enough to suggest measurement non-invariance. The range of the values for the effect size measures are presented in Table 11.3 for both language and gender. These measures represent small-to-moderate magnitude.

**TABLE 11.3 Ranges of effect size measures for language and gender DIF items**

| Groups | Domain | $\Delta R^2$ | $\Delta\beta$ |
|---|---|---|---|
| Language | Math | < .001–.044 | .001–.065 |
| | Reading | < .001–.031 | < .001–.045 |
| | Science | < .001–.018 | < .001–.015 |
| Gender | Math | < .001–.039 | .001–.018 |
| | Reading | < .001–.027 | <.001–.028 |
| | Science | < .001–.018 | .001–.010 |

The TCCs were either almost perfectly overlapped, for gender, or very close, for language. They are represented in Figures 11.2 and 11.3, respectively.

**FIGURE 11.2 Gender DIF detection: Test characteristic curves for mathematics items**

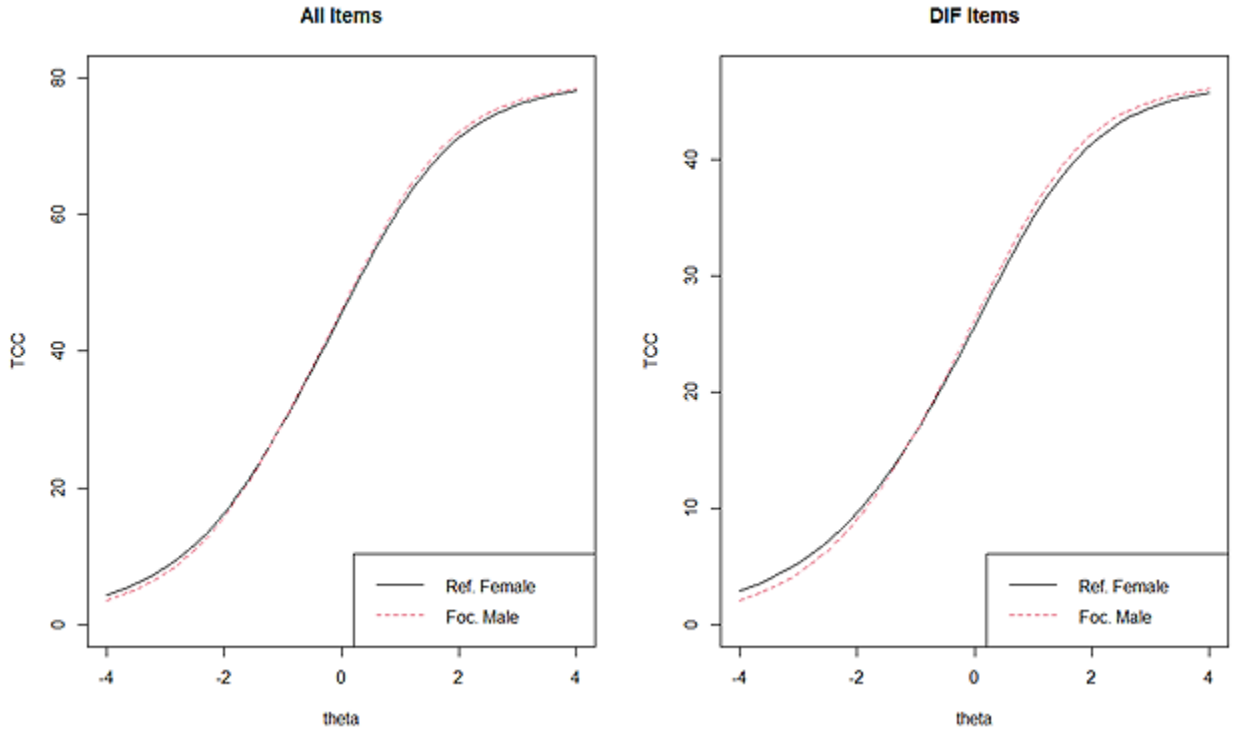
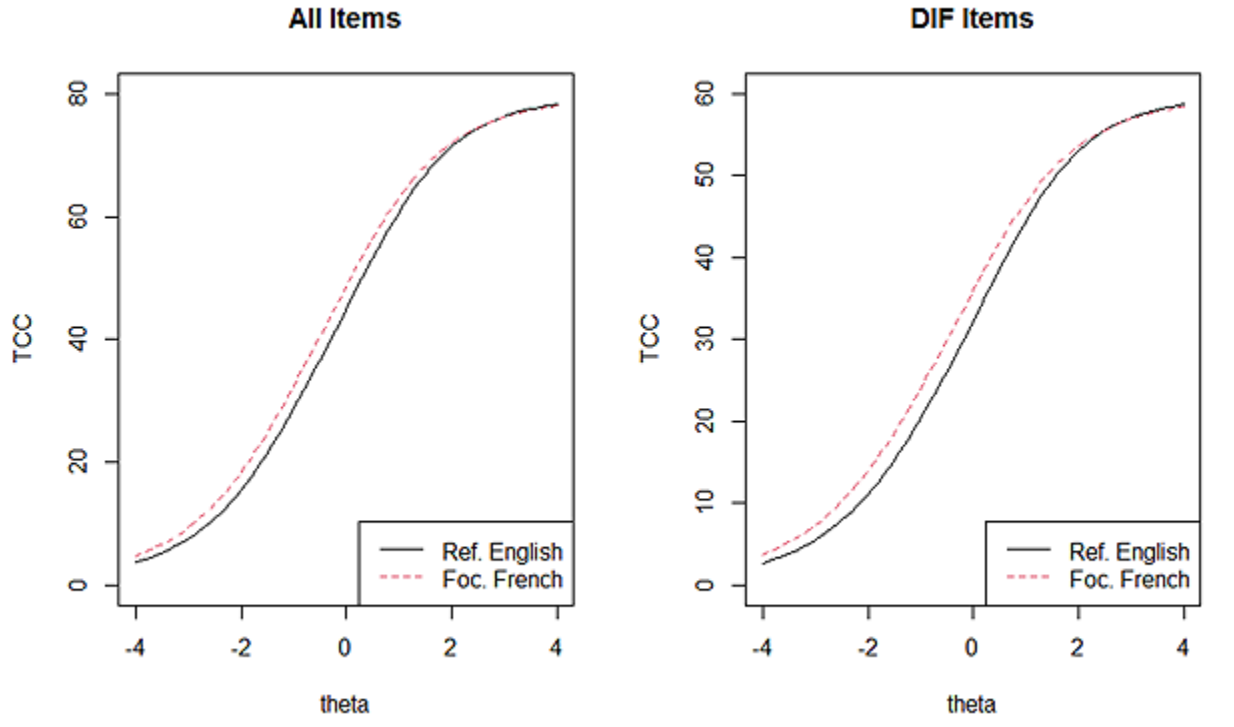



**FIGURE 11.3 Language DIF detection: Test characteristic curves for mathematics items**



## Test functioning

Test functioning was evaluated on the basis of the mean test score, the variability of the test scores, a measure (Cronbach's α) of internal consistency, the standard error of measurement, and the test information function.

## Linking and equating the minor domains with previous assessments

The linking and equating task provided a measure of the change from previous assessments to the current one. All three domains in PCAP 2019 included items that were used in previous assessments when these domains were the major one. No new items in the minor domains were developed for PCAP 2019. Therefore, all reading and science items were anchored. The design corresponded to the nonequivalent groups with anchor test (NEAT) design. With the change in the target population definition, 2010 was the baseline year for all the linking in the later cycles.[13] Given that no mode effect was found for the 2019 PCAP assessment, the linking of PCAP 2019 was done using concurrent calibration with PCAP 2016. Because the parameters of two successive assessment items were estimated simultaneously with anchor items common across years, the anchor item parameters had the same estimates and were on the same metric (de Ayala, 2009; Kim & Kolen, 2006). The approach had the advantage of making maximum use of all the available data in estimating item parameters (Martin et al., 2012).

With regard to theta scores, students from both samples were used to define the metric. Therefore, the proficiency score for the current assessment takers, when they were estimated using item parameters obtained under the concurrent calibration, were equated (de Ayala, 2009). However, the recalibration of the common items meant that their parameters were allowed to change over time. Because the parameters were allowed to vary over time, PCAP 2019 followed other large-scale assessment programs such as PIRLS and TIMSS that go a step further to incorporate this change into the linking process. More specifically, the PIRLS and TIMSS approach requires, once the concurrent calibration is performed, the following steps:

- estimating achievement distributions for the current assessment using the parameter from the concurrent calibration;
- determining the linear transformation that best matches the previous assessment's achievement distributions estimated under the concurrent calibration to the same assessment distributions obtained when the item parameters were estimated in the previous cycle; and
- applying the linear transformation determined in the preceding step (stage II) to the current assessment achievement distributions.

The generation of scores for all three domains for PCAP 2019 (the theta scores and the scale scores) used the item parameters estimated at this stage.

[13] In 2010, the comparison between 2007 and 2010 reading achievement was done using the 2010 item parameters as the baseline values (see CMEC, 2011). The decision to use 2010 as the baseline year instead of 2007 was made because of the shift of the targeted population from 13-year-old students to Grade 8/Secondary II students. Since 2010 became the baseline year, and to keep the comparison process consistent, the calibration therefore used the 2010 data sample for a reading trend measure.

## Achievement score generation and scale scores

For each student and in each of the three domains, the score generation occurred in three stages:

- A theta score was generated, reflecting the student's overall achievement in the domain of interest (i.e., mathematics, reading, or science), using the item parameters obtained from the concurrent calibrations. The estimation of the theta score used the Expected A Posteriori (EAP) method.
- The scores at stage II were weighted with the sampling weight on the scale with a Canadian mean of 500 and standard deviation of 100.
- Constants, determined through the linear transformation, were then applied to these weighted scores to generate the final scores that are directly comparable with achievement score from previous PCAP administrations.

It is important to note that this process also applied to the mathematics subdomains.

## Standard error estimates

The PCAP 2019 data analysis used a bootstrap approach in developing empirical standard error estimates for the Canadian results and means by province for each of the three achievement domains. The bootstrap method is one among the many types of replicate-based variance estimation techniques. Lohr (2010) discusses replication methods in detail. In short, these methods involved taking repeated subsamples, or replicates, from the data, re-computing the weighted survey estimate for each replicate, and the full sample, and then computing the variance as a function of the resulting estimates.

Initially, 1,000 replicate weights were created at the sampling stage. During the analytical step, it was determined that 500 would be sufficient for the precision of the estimates. In addition, using all 1,000 replicate weights would have been too computer-resource intensive and would have yielded precision estimates similar to a process using 500 replicate weights. (It is worth noting that the international large-scale assessments use fewer than 100 replicated weights.) Therefore, a total of 501 bootstrap replicate weights were used (*bsrwgt0–bsrwgt500*). The zeroth bootstrap weight (*bsrwgt0*), which is the same as the final student weight (*fsw*), was not used for calculating the standard error of a point estimate. On the other hand, the replicate weights *bsrwgt1–bsrwgt500* were used for this calculation. The inclusion of *bsrwgt0* follows the practice of many large-scale evaluation programs that implement this structure of replicate weights. It is worth noting that *bsrwgt0* and *fsw* can be used interchangeably to calculate a point estimate of interest.

## Presentation of the PCAP 2019 achievement results

Summary score reports were developed at the Canadian and provincial levels, and by language of the school system and gender, for each of the three achievement domains. Comparison of the achievement variations over time was also done for each of the three domains. Results were provided in tabular and graphic formats and followed the pattern set out in the PCAP 2007, 2010, 2013, and 2016 public reports. Standard errors were calculated using the bootstrap replicate weights. T-tests were conducted for all comparisons made, with the use of Bonferroni adjustments based on the number of comparisons where appropriate.

# 12 Analysis of Questionnaire Data

As in previous assessments, PCAP 2019 collected background data on students, teachers, and schools. The School Questionnaire was filled out by the school principals. The Student Questionnaire was completed online or on paper, depending on the students' mode of participation; for the Teacher and School Questionnaires, teachers and school principals completed their questionnaire online. The analysis of the questionnaire data included:

1. preliminary analysis
2. descriptive statistics
3. correlational analysis
    i. simple correlation
    ii. multiple linear regression modelling
4. principal component analysis to create derived variables where appropriate
5. item analysis for postulated and empirical constructed scales
6. group comparison analysis

These statistical analyses were conducted for each of the three questionnaires and were reported by language. PCAP 2019 is the second cycle in which teacher weights were included — in previous cycles, the school weights were used for teacher-level analyses, using the assumption that there was one teacher per selected class in each school. However, this was not the case in some schools, leading to inconsistency in the number of participating teachers and schools.

## Preliminary analysis

Preliminary analysis followed the same procedure used for the assessment items. It included data screening and recoding some items. Treatment of invalid data and missing values, however, differed slightly. Invalid responses (i.e., multiple responses to one question), omitted, and not-reached items were expected in the questionnaire data. They were all treated as missing values. However, not-administered items were not expected to appear in the data set because the full contextual questionnaire was administered to all students.

### *Data screening*

Data screening revealed that there were some cases in which some teachers and school principals filled out the questionnaire in both the online and paper formats, and there were inconsistencies in their responses. For these duplicate cases, efforts were made to determine the source of errors. For respondents where this could not be determined, the case with more responses was kept in the data sets. Frequency tables were produced for screening of each item:

- to check for anomalous data (e.g., outliers, errors);
- to examine the distribution of the response options (frequency and percentage); and
- to determine the missing rate per item and per booklet.

### *Item recoding*

The PCAP 2019 questionnaires included valid and invalid responses. A response to a question was valid if only one response option was chosen; the response was considered invalid if more than one option was chosen (with the exception outlined below). The task described here involved recoding raw valid and invalid responses to the items in the Student, Teacher, and School Questionnaires.

Invalid responses were coded 7 to distinguish them from valid and missing responses.

Some of the questionnaire items consisted of multiple-choice responses that allowed respondents to check all that apply, which required coding. Numeric or derived categorical variables were required for quantitative analyses of these items. These recoded items included the degrees or diplomas teachers held, the languages in which teachers taught mathematics, the types of feedback teachers gave to students in mathematics class, the grade levels taught in the school, and whether students identify themselves as Indigenous.

### *Missing data*

Three types of missing data occurred in the PCAP 2019 questionnaires:

- missing responses because a student ran out of time to complete the questionnaire (not-reached);[14]
- omitted items, that is, items skipped by a student intentionally or unintentionally throughout the instrument; and
- missing data because Teacher or School Questionnaires completed on paper were returned to CMEC after the data capture process was completed.

These types of missing data were coded 9. When it was possible, missing data were input using the multiple imputation (MI) procedure. Missing data present significant problems in statistical modelling because a case is typically deleted if missing data occur for any of the variables in the model. Even if only a few cases were missing for any one variable, the number of missing cases increases significantly if the missing data are scattered among the cases. Using techniques such as MI would alleviate the problem.

## Descriptive statistics

The descriptive statistics were produced by province, and by language of school systems, where appropriate, for background variables in the Student, Teacher, and School Questionnaires. They included percentage distributions for all items on categorical and Likert-type scales. The descriptive statistics also included the mean and the standard error.

[14] Teachers and principals are not restricted to assessment time limits, so missing data were not expected.

## Correlational analysis

The correlational analysis consisted of:

- computing simple correlation coefficients, also named the bivariate or the zero-order correlation, between student achievement and a background variable or an index variable. Correlational analysis results were then used to inform and confirm the decisions on the inclusion or exclusion of various variables in the subsequent analyses, which included linear multiple regression analysis, principal component analysis, item and indices analysis, and group comparison analysis
- performing linear multiple regression analysis to predict achievement in mathematics from a set of student-related variables
- performing linear multiple regression analysis to predict achievement at the class level, that is, the class mean achievement in mathematics from a set of teacher-related variables
- performing linear multiple regression analysis to predict achievement at the school level, that is, the school mean achievement in mathematics from a set of school-related variables

For all the correlation analysis, the dependent variable, student achievement, was assumed to be linearly related to the predictors. However, the linear regression assumptions were checked before conducting the analysis.

## Principal component analysis

In the questionnaires, sets or blocks of items were used to explore specific characteristics or attitudes. Principal component analysis (PCA) was used with the PCAP 2019 questionnaire items to reduce the complexity of the analysis and to obtain more stable measures.

As a first step, a list of potential blocks of items from the Student, Teacher, and School Questionnaires that might be considered for PCA was developed based on past contextual reports and the work of the PCAP questionnaire development group. Overall, 12 blocks of items from the Student Questionnaire were identified for PCA, 13 blocks of items from the Teacher Questionnaire, and 6 blocks of items from the School Questionnaire.

For each of the 31 individual PCAs, descriptive analysis (frequencies, mean, standard deviation), weighted and unweighted, was first performed to confirm the accuracy and validity of the data. In some cases, items needed to be recoded to match similar items for PCA before proceeding (e.g., Item 23 of the Teacher Questionnaire needed to be recoded in a series of eight dichotomous variables).

The following SPSS sample syntax was then used in each case. The VARIABLES option lists all the variables in the question, and the ANALYSIS option lists the variables retained for the analysis, which may change with each iteration as a result of deletions.

```
FACTOR
 /VARIABLES SCHQ18E SCHQ18F SCHQ18G SCHQ18H SCHQ18I SCHQ18J SCHQ18K SCHQ18L
SCHQ18M SCHQ18N SCHQ18O SCHQ18P SCHQ18Q
 /MISSING PAIRWISE
 /ANALYSIS SCHQ18E SCHQ18F SCHQ18G SCHQ18H SCHQ18I SCHQ18J SCHQ18L SCHQ18M SCHQ18N
SCHQ18O SCHQ18P SCHQ18Q
/PRINT UNIVARIATE INITIAL CORRELATION SIG KMO EXTRACTION ROTATION
 /FORMAT BLANK(.30)
/PLOT EIGEN
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION PC
/CRITERIA ITERATE(25) DELTA(0)
/ROTATION OBLIMIN
/SAVE REG(ALL)
/METHOD=CORRELATION.
```

The following criteria were then used to evaluate the output in order to identify variables for deletion:

- The correlation matrix: this provides bivariate correlations between all items. To prevent multicollinearity, bivariate correlations above 0.8 should be flagged, and one of the items should be removed from the analysis (the decision can be based on qualitative grounds).
- The Kaiser-Meyer-Olkin (KMO) and Bartlett's test: a value above the 0.5 threshold on the KMO test and a significance value at the 0.05 level on the Bartlett's test indicate that the items are suitable for PCA.
- The extraction communalities: variables with a small value of below 0.2 should be removed, as it indicates that the variables do not fit well with the component solution.
- Eigenvalue: a minimum Eigenvalue of 1 is the threshold for determining the number of components retained.
- Total variance explained: if the number of components with Eigenvalues above 1 cumulates less than 50 percent of the variance explained (some suggest going as high as 60 percent), it is recommended not to pursue the PCA further.
- Pattern matrix: when using an oblique (Oblimin) rotation, the pattern matrix represents the regression equation, where each standardized observed item is expressed as a function of the components. If a given item loads on more than one component, it should be deleted if its cross-loading difference is below 0.2 (e.g., the cross-loading value would be 0.05 for an item with loadings of 0.414 and 0.364 on two components, in which case the item should be deleted).

The last steps consist of checking the reliability of the scales identified in the PCA (see the SPSS sample syntax below). The VARIABLES option lists all the variables in the question, and the SCALE option includes only the variables retained in the analysis.

```
RELIABILITY
 /VARIABLES=STUQ27A STUQ27B STUQ27C STUQ27D STUQ27E STUQ27F STUQ27G
  STUQ27H STUQ27I STUQ27J STUQ27K STUQ27L STUQ27M STUQ27N STUQ27O STUQ27P
 /SCALE('Number') STUQ27A STUQ27D STUQ27E STUQ27J STUQ27L STUQ27M
 /MODEL=ALPHA
 /STATISTICS=DESCRIPTIVE SCALE
 /SUMMARY=TOTAL.
```

The following criteria were then used to assess the reliability of the identified scales:

- Cronbach's Alpha: the acceptable range of the Cronbach's Alpha statistics is dependent on the number of items, where Alpha values of 0.7 or above are expected for a scale (although this is difficult to use as an absolute criterion). Nonetheless, care should be exercised with Alpha values below 0.6.
- "Cronbach's Alpha if Item Deleted" column: this displays the Alpha value if a given item is deleted. If the value for an item is higher than the value of the Alpha for the scale, it signifies that the item is not contributing positively to the scale and, therefore, the item could be considered for deletion.

It should be noted that, if any item is removed as a result of the review at any of the steps listed above, a new analysis should be generated with the remaining items, thus repeating the same review process until only suitable items are retained.

Overall, 10 blocks of items from the Student Questionnaire were retained, resulting in the creation of 19 component scores; 12 blocks of items from the Teacher Questionnaire were retained, resulting in 20 component scores; and 6 block of items from the School Questionnaire were retained, resulting in 12 component scores.

The resulting component scores were then scaled to the mean of 50 and standard deviation of 10 to become index scores. For ease of interpretation, some scaled index scores such as that for the index of extracurricular activities (1 = "Yes", 2 = "No") were reverse-coded from the component scores so that a higher index score would indicate the presence of the extracurricular activities. In other instances, some of the component scores were reversed as a byproduct of the PCA process, and the scaled index scores were reverse-coded for the same reason.

Index scores were then correlated with mathematics achievement. Due to the large sample size, most correlations were found to be significant, although, in some cases, the correlation was low. As was the case in previous PCAP assessments, only components with correlations of .2 or above were reported in the PCAP 2019 contextual report. Index scores were also divided into four approximate equal quarters, weighted by student/teacher/school weights, as follows:

- bottom quarter: below the 25$^{th}$ percentile
- third quarter: 25$^{th}$ to 49$^{th}$ percentile
- second quarter: 50$^{th}$ to 74$^{th}$ percentile
- top quarter: 75$^{th}$ percentile or above

## Analyses of items and indices

Statistical analyses were performed on the items and the indices of the PCAP 2019 questionnaires. An analysis was conducted for each questionnaire and was reported by province, language of the school system, and gender. It primarily focused on the following items:

- mean and standard deviation
- correlations between items and with student achievement
- percentage distributions for all items within each index and the corresponding student achievement
- correlations and regressions between index scores and student achievement
- Cronbach's alpha for each index (see previous section)

## Group comparison analysis

The group comparison analysis involved:

- comparing student achievement means by student demographic variables;
- comparing student achievement means by the individual items of student, teacher, and school indices;
- comparing student achievement means by student, teacher, and school indices by quarters;
- comparing student index scores by gender and by the language of school systems;
- provincial comparisons of student index means with the Canadian means; and
- comparing student achievement means at the teacher (i.e., classroom) and school levels by teacher and school variables.

# 13 PCAP Data Sets

## Description of the data sets

All PCAP 2019 data sets are in English and French and are available to researchers. CMEC has several data sets for PCAP, including one covering all participating students, one covering all participating schools, and one covering teachers of the participating students. There is also a student/teacher/school data set containing all the student records merged with the questionnaire responses from teachers and school principals. This data set can establish relationships between student performance and the contextual data. The data sets come in SPSS, SAS, and Excel formats. Variables labels were set up beforehand on SPSS, so no codebook is provided for SPSS; however, a codebook and format file are provided for the data set in Excel and SAS format, respectively. To access the data set in SAS, run the SAS "Format" file before opening the data set.

### *Student data set*

This data set includes primarily the following data:

- general information about students (student, school, and class identification numbers; each student's province and language);
- student statistical weights;
- responses to the Student Questionnaire items;
- student index scores;[15] and
- achievement scores for all domains and for mathematics subdomains, and performance levels for mathematics.

This data set includes all the students with achievement scores but some cases may not contain questionnaire responses due to a lower number of students completing the questionnaire compared to the number of students who completed the assessment booklets.

### *Teacher data set*

This data set includes primarily:

- general information about teachers (teacher and school identification numbers; each teacher's province and language);
- teacher statistical weight;
- responses to the Teacher Questionnaire items; and
- teacher index scores.[16]

[15] Student index scores that were not included in the contextual report are also available in the data set.

[16] Teacher index scores that were not included in the contextual report are also available in the data set.

Because intact classes were used, one teacher was sampled in most schools, with two or more teachers sampled in a small number of schools.

### *School data set*

This data set includes primarily:

- general information about schools (school identification number, each school's province and language);
- school statistical weight;
- principals' responses to the School Questionnaire items; and
- school index scores.[17]

### *Merged data set – student/teacher/school*

This data set includes all the information from the student, teacher, and school data sets described earlier. This data set will enable researchers to establish relationships between student performance and the contextual data as reported by teachers and school principals. The number of cases in this data set is not equivalent to the number of cases in the student data set, as there are instances where two or more teachers were sampled in a small number of schools. Consequently, the responses of a single student needed to be replicated to match the responses of multiple teachers, as there would have been two or more teachers associated with the student. The student statistical weights and bootstrap replicate weights for these students are thus also adjusted accordingly.

Therefore, it should be noted that the merged data set is based only at the student level, and only the *adjusted* student weight and *adjusted* bootstrap replicate weights should be used for any analysis with the merged data set. As a result, any teacher or school characteristics obtained from the merged data set should be interpreted as follows: in the example of teachers' educational attainment, it should be reported as "the percentage of *students* whose teachers have a bachelor's degree," and *NOT* "the percentage of teachers with a bachelor's degree."

For analysis at the teacher and school levels, the teacher and school data sets should be used, respectively, along with the respective teacher and school statistical weights.

## Accessing the data set for research

PCAP, a pan-Canadian assessment with well-structured contextual questionnaires, affords unique opportunities for providing information related to key policy areas of concern to ministries and departments of education. PCAP allows provinces a simple way to compare their performance with that of the rest of Canada. PCAP data also provide information to provinces about the performance of their own education systems.

CMEC is committed to encouraging policy-relevant and educational research and maintaining, as a priority, the dissemination of research results to policy-makers and practitioners. The PCAP assessments were designed to yield achievement data at pan-Canadian and provincial/

[17] School index scores that were not included in the contextual report are also available in the data set.

territorial[18] levels. Data are also available by language of instruction, that is, English or French, and by gender. The sample size is too small, however, to yield reliable results from analysis within subcategories of a province (such as by schools or school boards/districts). For reasons of confidentiality, all information pertaining to the identity of students, schools, and school districts/boards is removed when final data sets are prepared for analysis by CMEC.

No data sets allowing for the identification of schools, school boards/districts, or individuals can be made available.

Researchers requesting access to the PCAP data sets will be asked to agree to the terms of availability described below.

## Terms and conditions

CMEC will maintain a registry of all requests for the use of PCAP data so that provinces/territories can be up to date about the research being undertaken using these data. Requests from researchers outside the field of education who are interested in using PCAP data are welcome.

For the purposes of the registry, researchers wishing to use PCAP data are asked to include the following information when requesting access to data sets:

- Name(s) and affiliation(s) of researchers working on the project (i.e., name of university, college, ministry/department of education, school district/board, research foundation, organization, etc., where the researcher is employed or for whom the researcher is undertaking the work)
- Contact information for the lead researcher on the project (mailing address, phone number, fax number, e-mail address)
- A succinct description of the project, including:
    - the purpose(s) of the project
    - the proposed methodology to be used for the research
    - the proposed sources of information and interviewees
    - CMEC documentation required to complete the research
    - the software to be used (to ensure compatibility with the PCAP database)
    - the proposed dissemination plan

Owing to sample-size considerations, researchers shall not use PCAP data to rank schools or school districts/boards, because such comparisons would not be valid.

Requests for access to confidential assessment materials such as test booklets will be considered by CMEC only with the strict assurance that booklet contents and identification numbers will not be divulged in any manner in the ensuing report.

Dissemination of results is a priority for PCAP research. CMEC is particularly interested in opportunities for dissemination to policy-makers and practitioners, and welcomes research initiatives

[18] No territories participated in PCAP 2019.

that include such activities. Publication of the research results will be the responsibility of the researcher(s), unless CMEC decides to play an active role in the dissemination of the research findings. The researcher(s) will be responsible for the research and its conclusions. The researcher(s) will be asked to submit a report of the research findings or a copy of the paper/journal article to CMEC prior to any publication or presentation of the findings. CMEC will distribute, under a confidentiality agreement, the report of the findings to member provinces/territories that are named or identified in any research findings one month prior to the publication or release of the findings, so that the province(s)/territory or territories involved can prepare communications strategies before the report is released. Unless otherwise agreed, this report would be used by CMEC for information purposes only, and CMEC would not publish the report without the consent of the researcher(s).

The source and original purpose for which the data were collected must be acknowledged when publishing or presenting secondary analysis of the data. The researcher(s) shall undertake to ensure that data sets are not made available to others by any means whatsoever.

## Information for researchers

CMEC is committed to encouraging policy-relevant research and prioritizing the dissemination of research results to policy-makers and practitioners. The *Assessment Matters!* series of policy-oriented research notes is designed to explore education issues in Canada, using results from national and international assessment programs. These research notes use relevant assessment data to answer pressing research questions about education issues in Canada.

Researchers may access, or request to access, data and other tools from international and pan-Canadian learning assessments on the CMEC website at https://www.cmec.ca/705/Learning_Assessment_Data_for_Researchers.html.

# References


Baker, F.B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147–162.

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1): 63–88.

Briggs, D. (2008, April). *An introduction to multidimensional IRT*. Paper presented at UC Berkeley. http://www.powershow.com/view/3c4039-MmRjY/An_Introduction_to_Multidimensional_IRT_Derek_Briggs_April_powerpoint_ppt_presentation

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309–329.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307–335.

Chen, W-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Choi, S.W., Gibbons, L.E., & Crane, P.K. (2011). *lordif:* An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1–30. http://www.jstatsoft.org/v39/i08/

Choi, S.W., Gibbons, L.E., & Crane, P.K. (2016a). *lordif: Logistic ordinal regression differential item functioning using IRT* (0.3-3). The R Project for Statistical Computing. https://cran.r-project.org/web/packages/lordif/index.html

Choi, S.W., Gibbons, L.E., & Crane, P.K. (2016b). *Package "lordif"* (0.3-3). The R project for statistical computing. https://cran.r-project.org/web/packages/lordif/lordif.pdf

Cizek, G.J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications Ltd.

Council of Ministers of Education, Canada (CMEC). (1997). *Common framework of science learning outcomes, K to 12: Pan-Canadian protocol for collaboration on school curriculum*. Author. http://science.cmec.ca/framework/

Council of Ministers of Education, Canada (CMEC). (2005). *The Pan-Canadian Assessment Program: Literature review of science assessment and test design*. Author (unpublished report).

Council of Ministers of Education, Canada (CMEC). (2011). *PCAP-2010: Pan-Canadian Assessment Program*. Author.

Council of Ministers of Education, Canada (CMEC). (2020). *PCAP 2019 assessment framework*. Author. https://cmec.ca/docs/pcap/pcap2019/PCAP-2019-Assessment-Framework-EN.pdf

Crane, P.K., van Belle, G., & Larson, E.B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*, 241–256.

de Ayala, R.J. (2009). *The theory and practice of item response theory.* Guilford Press.

de Ayala, R.J., Plake, B.S., & Impara, J.C. (2001). The impact of omitted response on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*(3), 213–234.

De Champlain, A.F., & Gessaroli, M.E. (1998). Assessing the dimensionality of item response matrices with small sample size and short test lengths. *Applied Measurement in Education, 11*, 231–253.

Fensham, P., & Harlen, W. (1999). School science and public understanding of science. *International Journal of Science Education, 21*(7), 755–63.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225–245.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimension and allocating items. *Journal of Educational Measurement, 42*, 149–169.

Foy, P., Brossman, B., & Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 achievement data. In M.O. Martin & I.V.S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/methods/pdf/TP11_Scaling_Achievement.pdf

Fraser, C., & McDonald, R.P. (2003). NOHARM: *A window program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [Computer program]. Niagara College. http://noharm.software.informer.com/

Fu, J., & Monfils, L. (2016). *LRDIF_ES: A SAS macro for logistic regression tests for differential item functioning of dichotomous and polytomous items*, Research Memorandum No. RM-16-17. Educational Testing Service.

Hidi, S., & Berndorff, D. (1998). Situational interest and learning. In L. Hoffmann, A. Krapp, K.A. Renniger, & J. Baumert (Eds.), *Interest and learning*. Institute for Science Education at the University of Kiel.

Hoy, A.W. (2000, April). *Changes in teacher efficacy during the early years of teaching*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329–349.

Johnson, M.S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software, 20*(10), 1–19.

Kandel, L., & Moles, A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers Études de Radio-Télévision, 19*, 253–274.

Kim, S., & Kolen, M.J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Psychological Measurement, 19*(4), 357–381.

Klare, G.R. (1988). The formative years. In B.L. Zakaluk & S.J. Samuels (Eds.), *Readability, its past, present and future*. International Reading Association, 14–34.

Knol, W.R., & Berger, M.P.F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457–477.

Laveault, D., & Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en education* (2nd ed.). De Boeck.

Lohr, S. (2010). *Sampling: Design and analysis* (2nd ed.). Duxbury Press.

Lord, F.M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95–100.

Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Erlbaum.

Lord, F.M. (1983). Maximum likelihood estimation of item parameters when some responses are omitted. *Psychometrika, 48*, 477–482.

Ludlow, L.H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59*, 615–630.

Maldonado, G., & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology, 138*(11), 923–936.

Martin, M.O., Mullis, I.V.S, Foy, P., Brossman, B., & Stanco, G.M. (2012). Estimating linking error in PIRLS. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 5, 35–47. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_05_Chapter_2.pdf

McDonald, R.P. (1967). *Nonlinear factor analysis*. Psychometric Monographs, No. 15. Psychometric Corporation. http://www.psychometrika.org/journal/online/MN15.pdf

Muraki, E., & Engelhard, G. (1989, April). *Examining differential item functioning with BIMAIN*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Author.

O'Grady, K., Fung, K., Servage, L., & Khan, G. (2018). *PCAP 2016: Report on the pan-Canadian assessment of reading, mathematics, and science*. Council of Ministers of Education, Canada. https://cmec.ca/Publications/Lists/Publications/Attachments/381/PCAP-2016-Public-Report-EN.pdf

O'Grady, K., Houme, K., Costa, E., Rostamian, A., & Tao, Y. (2021). *PCAP 2019: Report on the pan-Canadian assessment of mathematics, reading, and science*. Council of Ministers of Education, Canada. https://cmec.ca/Publications/Lists/Publications/Attachments/426/PCAP2019-Public-Report-EN.pdf

Organisation for Economic Co-operation and Development (OECD). (2006). *PISA 2006: Science competencies for tomorrow's world*. PISA, OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2012). *PISA 2009 technical report*. PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education, 25*(9), 1049–1079.

Riggs, I., & Enochs, L. (1990). Towards the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education 74*, 625–637.

Statistics Canada. (2020). English, French and non-official languages spoken at home by geography, 2001 to 2016. Table 15-10-0009-01. https://doi.org/10.25318/1510000901-eng

Vector Psychometric Group. (2020, June). *IRTPRO guide*. https://vpgcentral.com/wp-content/uploads/2020/06/IRTPROGuide.pdf

Wang, M.C., Haertel, G.D., & Walberg, H.J. (1990). What influences learning? A content analysis of review literature. *Journal of Educational Research, 84*(1), 30–43.

Wang, M.C., Haertel, G.D., & Walberg, H.J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63,*(3), 249–294.

Wang, M.C., Haertel, G.D., & Walberg, H.J. (1994). Synthesis of research: What helps students learn? *Educational Leadership*, December 1993/January 1994, 74–79.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Zhang, B., & Walter, C.M. (2008). Impact of missing data on person-model fit and person trait estimation. *Applied Psychological Measurement, 32*(6), 466–479.

# APPENDIX A: SCHOOL AND CLASS SAMPLE DESIGN

## 1. Introduction

The Council of Ministers of Education Canada (CMEC) was responsible for administering the fifth iteration of the Pan-Canadian Assessment Program (PCAP). PCAP, which took place in the spring of 2019, is designed to evaluate the competencies of Grade 8 (Secondary II in Quebec) students in the fields of reading, mathematics, and science.

CMEC contracted Statistics Canada to design and implement the probability school sampling strategy for PCAP 2019. The purpose of this appendix is to detail all aspects of this sample design. More specifically, this appendix covers such topics as the target and survey populations, the survey frames, stratification, the sample sizes and sample selection methods, as well as the proposed methodology and schedule of sampling activities.[19]

## 2. Sample design

*Sample design* refers to what a sample consists of and how it is to be obtained. The sample design is a set of specifications that describe the target and survey populations, the survey frame, the stratification, the sample size, and the sample selection methods. These design features are described in detail in sub-sections 2.1 to 2.5.

As mentioned above, the objective of PCAP is to assess certain competencies of Grade 8/Secondary II students from Canada's ten provinces. Since a single list of all Grade 8/Secondary II students in Canada does not exist, the PCAP sample will be selected utilizing a stratified two-stage probability sampling design. This means that the sample of Grade 8/Secondary II students is selected in two successive stages.

In the first stage, a random sample of schools (primary sampling units – PSUs) is selected from a list of in-scope schools provided by CMEC. All in-scope Grade 8/Secondary II classes within the selected schools are then enumerated. The second stage of sample selection consists of randomly selecting one Grade 8/Secondary II class per selected school. The Grade 8/Secondary II classes are, therefore, the secondary sampling units (SSUs). All in-scope students within the selected Grade 8/Secondary II classes receive the PCAP evaluation.

The 2019 PCAP assessment was online for the first time. To assess the impact of changing the assessment from a paper mode to an online mode, an extra sample of schools in large non-census strata was selected. The students in this sample were assessed using the paper-based assessment.

### 2.1 Target and survey populations

Defining the population from which a sample is selected is an essential step in developing a sound sample design. A good definition facilitates the sampling process and prevents ambiguities. Populations can be referred to as either target or survey populations.

[19] This document was prepared by the Statistical Consultation Group, International Collaboration and Methodology Innovation Centre (ICMIC) at Statistics Canada and the Council of Ministers of Education, Canada.

The *target population* is the set of elements about which information is wanted and estimates are required. The *survey population* is a subset of the target population that arises from the survey design and other practical considerations. The survey population may not be exactly the same as the target population; however, ideally, it should be very similar.

The target population for this study was all Grade 8/Secondary II students in the ten Canadian provinces. Schools that were not funded by the provinces were excluded from the target population. The following types of students were also excluded from the target population:

- students with functional or intellectual disabilities
- students from federal or international schools
- students that have been in Canada for less than two years and who speak neither English nor French

For practical reasons, the survey population for this study excluded certain schools with certain characteristics and schools with fewer than five Grade 8/Secondary II students. Our analysis suggests that these exclusions account for no more than 1.7 percent.

Two types of units generally have to be distinguished in the survey population: the sampling units and the respondent units. For this study, there are two types of sampling units. At the first stage, the sampling units are schools. At the second stage, the sampling units are classes of Grade 8/Secondary II students. For this study, the responding units are the Grade 8/Secondary II students within the selected classes.

The reference period is defined as the time period to which the data refer. For this study, the reference period is defined as the time period during which the PCAP evaluations take place. The most recent one took place in the spring of 2019.

## 2.2 Survey frame

A *survey frame* is any list, material, or device that delimits, identifies, and allows access to elements of the survey population. The frame contains all the units that comprise the population from which the sample is drawn.

The quality of the frame determines the coverage of the survey and influences the efficiency of the survey design. The erroneous omission or inclusion of units on the frame will respectively lead to under-coverage and over-coverage in the study.

For this study, there were two survey frames — one for each stage of sample selection. Sub-section 2.2.1 gives further detail on the survey frame for the first stage of sample selection.

### *2.2.1 First stage survey frame (list of in-scope schools)*

For the first stage of sample selection, CMEC obtained a list of all in-scope schools from the provinces. This list was provided to Statistics Canada for use as the first stage survey frame. This list contained a school identification number, province, language of the school board (either English or French), the number of Grade 8/Secondary II students within the school, the number of Grade 8/Secondary II classes, and an exclusion flag indicating whether or not the school should be excluded from the survey coverage as well as the reason for exclusion (Table A.1).

### Table A.1 Important variables for the first stage of sample selection

| Variable | Description |
|---|---|
| *SCHOOLID* | School identification number |
| *PROVINCE* | Province where the school is located |
| *LANGUAGEID* | Language of school board (English or French) |
| *G8_STUDENTS* | Number of Grade 8/Secondary II students |
| *G8_CLASSES* | Number of Grade 8/Secondary II classes |
| *EXEMPTION_CODE* | School exclusion flag |

Section 2.1 delineated the few types of schools that are not covered by this study. In order to remove these schools from the survey frame before sampling, the exemption flag was utilized. Table A.2 provides documentation for the types of schools excluded from this study.

### Table A.2 School exclusions by type

| Exemption | Schools | Students | Percent |
|---|---|---|---|
| Very small school (fewer than 5 students) | 633 | 1,423 | 0.4 |
| Special needs school | 58 | 1,329 | 0.4 |
| School within other provinces | 3 | 387 | 0.1 |
| Geographically isolated school | 13 | 68 | 0.0 |
| International school; offshore | 4 | 440 | 0.1 |
| School that is not funded | 1 | 351 | 0.1 |
| Hospital school | 5 | 139 | 0.4 |
| Youth school | 27 | 1,525 | 0.0 |
| District school board with special status | 30 | 750 | 0.2 |
| Total | 774 | 6,412 | 1.7 |

In order to verify the coverage of the PCAP 2019 frame, Statistics Canada compared the number of Grade 8/Secondary II students reported on the frame to census population projections of 13-year-olds. These counts were also compared to those observed on the previous two PCAP cycles frames (Table A.3). Considering these results, Statistics Canada ensures that the number of Grade 8/Secondary II students on the 2017 frame aligns well with the counts from the previous two cycles' frames as well as with the census projections. In 2017, at the national level, Statistics Canada ensured that the PCAP frame offered good population coverage, with the frame counts matching almost exactly to the population projections. At the provincial level, the population coverage is on average in the mid-nineties. However, this figure is as low as 89.99 percent in Manitoba. Given these comparisons, one has confidence in the quality of the PCAP frame.

### TABLE A.3 Coverage of the PCAP frame

| | Population Projections* (13-year-olds) | | | Frame** | | | Proportion of Population (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2013 | 2015 | 2017 | 2013 | 2015 | 2017 | 2013 | 2015 | 2017 |
| NL | 5,436 | 5,221 | 5,293 | 5,441 | 5,193 | 5,048 | 100.1 | 99.5 | 95.37 |
| PE | 1,657 | 1,527 | 1,664 | 1,487 | 1,433 | 1,703 | 89.7 | 93.8 | 102.34 |
| NS | 9,785 | 9,146 | 9,046 | 9,792 | 8,674 | 8,818 | 100.1 | 94.8 | 97.48 |
| NB | 8,010 | 7,543 | 7,683 | 7,976 | 7,327 | 7,269 | 99.6 | 97.1 | 94.61 |
| QC | 79,921 | 78,354 | 79,848 | 85,435 | 81,563 | 83,196 | 106.9 | 104.1 | 104.19 |
| ON | 153,091 | 147,322 | 151,152 | 145,756 | 146,765 | 150,318 | 95.2 | 99.6 | 99.45 |
| MB | 16,289 | 15,852 | 16,137 | 14,451 | 14,174 | 14,522 | 88.7 | 89.4 | 89.99 |
| SK | 13,677 | 13,371 | 14,102 | 12,598 | 12,335 | 12,935 | 92.1 | 92.3 | 91.72 |
| AB | 45,386 | 45,354 | 48,118 | 40,094 | 46,578 | 48,599 | 88.3 | 102.7 | 101.00 |
| BC | 47,129 | 45,848 | 46,881 | 46,895 | 45,576 | 45,627 | 99.5 | 99.4 | 97.33 |
| **Canada** | **380,381** | **369,538** | **379,924** | **369,925** | **369,618** | **378,035** | **97.3** | **100.0** | **99.50** |

*Source: Statistics Canada, Table 17-10-0005-01, with data derived from estimates of population, by age group and sex, for July 1, Canada, provinces and territories, annual (persons unless otherwise noted)
** Frame was produced based on the list of schools provided by the provinces to CMEC

## *2.2.2 Second stage survey frame (list of in-scope classes)*

Once a sample of schools is selected, CMEC works with the provinces to enumerate all Grade 8/Secondary II classes within the selected schools. As Grade 8/Secondary II students may move between classes throughout the day, it is recommended that the enumeration be of Grade 8/Secondary II homerooms or some class that every Grade 8/Secondary II student has to take (e.g., math or mother tongue). Table A.4 documents the important variables required for the second stage of sample selection.

### TABLE A.4 Important variables for the second stage of sample selection

| Variable | Description |
|---|---|
| *CLASS ID* | Within each school, a unique class identification number |
| *TEACHER NAME* | Name of mathematics teacher |
| *GRADE 8 /SECONDARY II STUDENTS* | Number of Grade 8/Secondary II students within each class |
| *EXEMPTION* | Class exemption |

Section 2.1 delineated the students who are not covered by this study. These students, or classes of students, were removed from the frame before sampling. An entire class can be exempted if all the students are in a category for which students are exempted; these three categories, and associated category codes, are given below.

F *Exempted because of functional disabilities:* A student who has a physical disability and who is unable to perform in the PCAP testing situation, even with one or more of the seven permitted accommodations, should be exempted. A student who has a functional disability but is, nevertheless, able to participate should be included in the testing.

I *Exempted because of intellectual disabilities or socio-emotional conditions:* A student who, in the professional opinion of the school principal or other qualified staff members, is considered to have an intellectual disability, or a socio-emotional condition, or has been tested as such, should be exempted. This category includes students who are emotionally or mentally unable to follow even the general instructions for the test.

N *Exempted because of language (non-native speakers):* This exemption is applicable only to those students whose mother tongue is neither English nor French. In large-scale assessments, schools can consider such students who have been in Canada for less than two years to be exempt.

All exemptions at the class level were approved by CMEC. Provinces were also asked to provide explanations for why certain sampled schools were not participating in PCAP.

## 2.3 Stratification

*Stratification* is a means of organizing the sampling frame so that better precision can be achieved with a fixed sample size. Stratification can also be used to guarantee that a minimum number of units or precision requirements for certain population groups will be obtained. Strata are exhaustive and are mutually exclusive groups of schools. Each school is in one and only one stratum. The total sample size is separated among the strata. Each stratum is sampled independently.

CMEC expressed a need to publish reliable statistics at the national and provincial levels, as well as by the language of the school boards/districts within provinces. In order to ensure a large enough sample within these domains, the PCAP strata were defined as the cross-classification of the province by the language of the school board/district. In the province of Quebec, in addition to stratifying by language, schools were stratified by their type, public or private.

## 2.4 Sample size allocation

*Sample size allocation* is a process of compromise in which the precision requirements of the estimates are weighed against various operational constraints such as time, budget, and available resources.

For the 2019 PCAP, Statistics Canada started the sample size allocation process by considering the allocation strategy that was used in 2016. Provinces requested that the same precision achieved in 2016 be maintained in 2019. Statistics Canada therefore used the same allocation scheme in all strata except Saskatchewan – anglophone. The allocation in Saskatchewan – anglophone was lowered to 158 schools to align with the other large strata. In addition to the CMEC request to achieve the same precision, this allocation had the further advantage of being pre-approved by the provinces. An extra sample of 158 schools was selected, with their students to be assessed using the paper-based assessment. The sample size was determined so that the overall standard error does not exceed 3.5. This sample was allocated proportionally among the large non-census strata. This was requested by CMEC to keep the census strata as censuses. Because the paper sample does not represent all the provinces, the mode study results should be applied only to the strata where a paper sample was selected. Table A.5 contains the sample allocation per stratum. The table contains the allocation for the online and paper samples.

### Table A.5 The 2019 PCAP school sample size allocation

| Stratum | In-scope population | | Online sample | | Paper sample |
|---|---|---|---|---|---|
| | Schools | Students | Schools | Census of schools | Schools |
| Alberta – anglophone | 756 | 47,226 | 152 | | 20 |
| Alberta – francophone | 18 | 428 | 18 | Yes | |
| British Columbia – anglophone | 424 | 45,171 | 150 | | 20 |
| British Columbia – francophone | 13 | 289 | 13 | Yes | |
| Manitoba – anglophone | 320 | 13,753 | 153 | | 8 |
| Manitoba – francophone | 16 | 366 | 16 | Yes | |
| New Brunswick – anglophone | 83 | 5,127 | 83 | Yes | |
| New Brunswick – francophone | 60 | 2,131 | 60 | Yes | |
| Newfoundland – anglophone | 112 | 4,930 | 112 | Yes | |
| Newfoundland – francophone | 1 | 8 | 1 | Yes | |
| Nova Scotia – anglophone | 114 | 8,478 | 114 | Yes | |
| Nova Scotia – francophone | 11 | 339 | 11 | Yes | |
| Ontario – anglophone | 3,105 | 140,943 | 151 | | 62 |
| Ontario – francophone | 176 | 7,414 | 126 | | 8 |
| Ontario – empty* | 318 | 1,590 | 4 | | |
| Prince Edward Island – anglophone | 22 | 1,639 | 22 | Yes | |
| Prince Edward Island – francophone | 4 | 64 | 4 | Yes | |
| Quebec – anglophone – public | 67 | 6,420 | 67 | Yes | |
| Quebec – anglophone – private | 24 | 1,168 | 24 | Yes | |
| Quebec – francophone – public | 336 | 56,522 | 120 | | 24 |
| Quebec – francophone – private | 115 | 14,992 | 31 | | 8 |
| Saskatchewan – anglophone | 466 | 12,545 | 150 | | 8 |
| Saskatchewan – francophone | 6 | 80 | 6 | Yes | |
| **Canada** | **6,567** | **371,623** | **1,588** | | **158** |

* This stratum contains schools that were empty in the frame but had the possibility of having Grade 8 students in 2019. All schools in this stratum were in the anglophone school sector.
*Note:* Although students in francophone schools in Prince Edward Island and Newfoundland and Labrador were sampled, the results were not reported due to small sample size.

## 2.5 School and class sample selection

*Sample selection* refers to the process used to obtain the survey sample from the survey population. The purpose of this sub-section is to document the sample selection process for both the first and second stages of sample selection.

For the first stage of sample selection (the selection of schools), two methods were used. For 14 of 21 strata, a census of schools was taken. For these strata, the sample selection method was straightforward, as all schools were selected into the sample. For the remaining seven strata, the sample selection method was Systematic Sampling (SYS), where the schools were first sorted in descending order by the number of Grade 8/Secondary II students within each school. This sorting is a simple way of ensuring that the sample of schools is allocated proportional to the size of the schools. The advantage of this method is twofold. First, it is administratively convenient, as replacement schools are taken to be the schools directly above and below the selected schools on the sorted list of schools. The

second advantage of this method of sample selection is that it utilizes available auxiliary information at the sample selection stage. As a result, if the estimates of interest are correlated to the size of the schools, then the resulting estimates may have less sampling variability than a sampling scheme that did not incorporate this auxiliary information. The disadvantage of this method is that if the school size is not correlated to the estimates of interest, then the resulting estimates may have more sampling variability than a sampling scheme that ignores this auxiliary data.

CMEC made every effort to confirm the participation of as many sampled schools as possible. This was important in order to minimize the potential for non-response biases. After all contacts with sampled schools were made, CMEC contacted replacement schools for those sampled schools that did not participate. Each sampled school that did not participate was replaced, if possible, by the first replacement school. Second replacement schools were used only if both the corresponding sampled school and first replacement school did not participate. *If the original sampled school was ineligible, or was a type of school belonging to defined school-level exclusions or was closed, then no replacement schools was used.* Additionally, a school with a small but sufficient number of Grade 8/Secondary II students was not replaced simply because the number of students was smaller than expected.

In general, replacement schools are used as a treatment for non-response (i.e., if and only if the original school refused to participate). In survey methodology, non-response is dealt with in one of two ways. The first is with a weight adjustment. This is where respondents are allowed to represent the non-respondents. The second way is known as imputation. Here, response values are assigned to non-respondents using data collected from respondents. Under this framework, replacement schools are viewed as a form of imputation. In the case when out-of-scope schools are chosen, Statistics Canada do not wish to assign this weight to other schools or treat them with imputation, as these selected out-of-scope schools effectively represent all of the other out-of-scope schools on the frame.

In short, replacement schools are used to treat non-response. For 2019, 26 schools out of 1,429 respondent schools were replacement schools. Out-of-scope schools receive no treatment, as they represent the fact that the frame is out-of-date. This lost sample size is the price one pays for having an out-of-date frame.

For the second stage of sample selection (i.e., the selection of classes), among those schools that had more than 20 students and were able to enumerate all Grade 8/Secondary II classes, one Grade 8/Secondary II class per school was selected via a Simple Random Sample (SRS). All in-scope students within the selected classes were assessed. For schools that had more than 20 students but were unable to enumerate Grade 8/Secondary II classes, an enumeration of all Grade 8/Secondary II students was completed, and a SRS of 20 students was selected. For schools with 20 or fewer Grade 8/Secondary II students, a census of students within these schools was taken. Once the list of classes was selected, the total number of required students in each stratum was compared with the total number of students in the sampled classes. If the number of students sampled from the classes did not meet the expected numbers required as indicated by Statistics Canada, a second class in the bigger schools was selected.

### 2.5.1 Minimum overlap with TIMSS 2019

Minimum overlap was performed with schools selected to participate in TIMSS 2019. The method used is a technique described in Chowdhury, Chu, and Kaufman.[20] This method adjusts the probability of selecting a school to participate in PCAP 2019, depending on whether the school was selected to participate in TIMSS 2019 or not.

## 3. Sampling weights

Sub-sections 2.5 and 2.5.1 described, in detail, the first and second stage sample selection methods. The purpose of this section is to specify the design weights, which arise from the above-mentioned sample selection methods.

Note that weighting and adjustments were done independently for the electronic and paper samples. In other words, both samples were weighted to represent the population they were selected from. The electronic sample with a sample size of 1,588 schools will give greater precision than the paper sample of 158 schools. The formula described below can be used for both samples. A set of weights was actually delivered for both samples. The purpose of the paper sample was to evaluate if there was a collection mode effect in going from a paper-based test to a computer-based test for data collection.

The first stage of sample selection (the selection of schools) results in each school being selected with some inclusion probability. The inverse of this inclusion probability is the school's design weight. In order to define the design weights, some notation is required. Assume the set of schools in the survey population is of size $N$. Let $h$ represent the strata of which $H = 23$. Let $n_h$ be the sample size allocated to stratum $h$. Let $X_i$ be the total number of Grade 8/Secondary II students in school $i$. Then the design weight ($d_i^{school}$) for the selected school $i$ is given as:

$$d_i^{school} = \begin{cases} \dfrac{\Sigma_{i \varepsilon h} X_i}{n_h X_i}, & if\ n_h < N_h \\ 1, & if\ n_h = N_h \end{cases}$$

To account for the second stage of sample selection (selecting classes from within schools), we apply the following weight adjustment ($adj_i^{class}$) to the above school weight:

$$adj_i^{class} = \begin{cases} m_i, & if\ X_i > 20\ and\ all\ classes\ are\ enumerated \\ \dfrac{X_i}{20}, & if\ X_i > 20\ and\ students\ are\ enumerated \\ 1, & if\ X_i \leq 20 \end{cases}$$

where $m_i$ is the total number of Grade 8/Secondary II classes within school $i$. The student design weight ($d_i^{student}$) is defined as:

$$d_i^{student} = d_i^{school} * adj_i^{class}$$

[20] S. Chowdhury, A. Chu, and S. Kaufman, "Minimizing Overlap in NCES Surveys," *Proceedings of the Survey Methods Research Section, American Statistical Association* (2000), 174–179.

In 2019, out of the 1,429 respondent schools, 594 schools had more than 20 Grade 8/Secondary II students and all classes were enumerated, and 835 schools had 20 Grade 8/Secondary II students or less.

To arrive at the estimation weights — those weights to be used in the production of the final estimates — further weight adjustments for school non-response and student non-response were applied. In order to facilitate the calculation of the student-level non-response weight adjustments, it was necessary for CMEC to provide Statistics Canada's with the total number of in-scope students per selected class, or at the school if classes could not be enumerated.

CMEC, in collaboration with the schools and the provinces, conducts PCAP evaluations. For 2019, mathematics was the major domain, and reading and science were the two minor domains. Once again, all selected students or all in-scope students within selected classes were evaluated. As there are four distinct test booklets, the distribution of the booklets, per Statistics Canada's recommendation, had a single random starting point within each province. As well, it was important that the distribution of booklets be coordinated across schools. That is, School 1 started distributing booklets at the random start, and School 2 started distributing booklets where School 1 had left off, and so on. This method ensured equitable distribution of each booklet.

## 3.1 Sampling weights non-response adjustments

School weights for the participating schools were inflated in order to account for the non-responding schools. Schools that were exempt retain their weight, representing all the other exempt schools that exist in the population. In 2019, out of 1,588 schools, 1,429 responded, and 159 were non-respondents.

In practice, the school non-response adjustment occurred in two stages. There is the phase where classes had been selected and schools could decide not to participate, referred to as initial non-response. There is also the case where, subsequently, after the data had been returned, additional schools were found to be non-participating.

Recall the initial school weight for school $i$ in stratum $h$ ($d_{hi}^{school}$, {**School_design_weight**}[21]):

$$d_{hi}^{school} = \begin{cases} \dfrac{\Sigma_{ieh} X_i}{n_h X_i}, & if\ n_h < N_h \\ 1, & if\ n_h = N_h \end{cases}$$

then the school non-response adjustment factor for each participating school in *stratum h* ($nr\_adj_h^{school}${**school_nr_adjust**}):

$$nr\_adj_h^{school} = \frac{\Sigma_i I_{hi}^{p\_school} + \Sigma_i I_{hi}^{nr-school}}{\Sigma_i I_{hi}^{p\_school}}$$

where:

[21] The corresponding variable names on the files are given in { } for the electronic sample. For the paper sample, replace "_*e*" with "_*p*" in the variable name.

$$I_{hi}^{p\_school} = \begin{cases} 0 \text{ if school is not participating} \\ 0 \text{ if school is out of scope} \\ 1 \text{ if school is participating} \end{cases}$$

$$I_{hi}^{nr-school} = \begin{cases} 1 \text{ if school is not participating} \\ 0 \text{ if school is out of scope} \\ 0 \text{ if school is participating} \end{cases}$$

with the final school weight ($w_{final_i}^{school}$ {**final_weight_schools_e**}) given by:

$$w_{final_i}^{school} = nr\_adj_h^{school} * d_i^{school}$$

We calculated class adjustments ($adj_i^{class}$) for the participating schools. The next step was to derive student weights. An estimation weight for each student consists of the final school weight times the class adjustment times the student weight adjustment to account for non-response.

Student weight adjustments are created similarly to the school weight adjustments; students who did not respond due to absence, lack of permission to write, or the answer sheet/booklet not being returned are all considered non-respondents.[22] This relies on the missing at random assumption — i.e., that the students who did not participate are similar to those who did. The student non-response adjustment ($nr\_adj_i^{student}${**student_nr_adj**}), for school $i$ in the selected class is expressed as:[23]

$$nr\_adj_i^{student} = \frac{\#\ participating\ students\ in\ the\ class + \#\ non-respondent\ students\ in\ the\ class}{\#\ participating\ students\ in\ the\ class}$$

Using a concrete example, if there were 20 students in the class, and 10 participated in the assessment, 2 were absent, and 8 were exempted for various reasons, then the 10 who did participate would have a non-response adjustment of 1.2 (=12/10 = (10+2)/10).

Therefore, the final analysis weight for each student ($w_j^{*student}$ {**student_weight_final** }) for student $j$ in school $i$, in stratum $h$ is given by:

$$w_j^{student} = d_i^{school} * nr\_adj_h^{school} * adj_i^{class} * nr\_adj_i^{student}$$

At the school level, the final analysis weight is given by:

$$w_{final_i}^{school} = nr\_adj_h^{school} * d_i^{school}$$

or using the variable names:

**student_weight_final**

**= School_design_weight * school_nr_adjust * class_adj * student_nr_adj**

**school_weight_final**

**= School_design_weight * school_nr_adjust**

[22] See Table A.7 for a list of participation status and associated response codes.

[23] "Class" being either an actual class, or a grouping of (small) classes to make a *de facto* class.

## 3.2 Class-level weights and adjustments

Class-level data were collected by interviewing all mathematics teachers for each selected class. For some classes, more than one teacher per class was interviewed. This means that there was more than one observation by class for classes with more than one teacher — i.e., the number of observations by class is equal to the number of math teachers surveyed for this class. The class-level weights were obtained by multiplying the school final weight by the class adjustment and by an adjustment for the number of observations (teachers). At the class level, the final analysis weight is given by:

$$w_{final_i}^{class} = w_{final_i}^{school} * adj_i^{class} * tch\_adj_i$$

where:

$$tch\_adj_i = \frac{1}{\#\ teachers\ surveyed}$$

or using the variable names:

**class_weight_final**
**= School_final_weight * class_adj * tchr_adj**

## 4. Bootstrap weights

The bootstrap method belongs to a family of replicate-based variance estimation techniques. A detailed discussion of replication methods can be found in Lohr.[24] Such methods involve the taking of repeated subsamples, or replicates, from the data, re-computing the weighted survey estimate for each replicate and the full sample, and then computing the variance as a function of the resulting estimates.

In order to allow for analysts to create estimates of variability that properly account for the complex design of PCAP, bootstrap weights were created for only the participating students.

The method of Rao and Wu[25] for the estimation of the bootstrap weights was used, and, given the large sampling fractions in many of the strata, the variant for the bootstrap weights as presented in Beaumont and Patak[26] was used, with the bootstrap weight adjustment ($a_k$) given by:

$$a_{h,k} = 1 - \sqrt{(1-f_h)} + \frac{n_h}{n_h - 1}\sqrt{(1-f_h)}m_k^*$$

where:

- $n_h$ is the number of participating schools in stratum h
- $f_h$ is the ratio of the participating schools to the total number of schools in stratum h
- $m_k^*$ is the number of times school $k$ was chosen out of the *n-1* trials

[24] S. Lohr, *Sampling*: *Design and Analysis*, 2nd ed. (Duxbury Press, 2010).

[25] J.N.K. Rao and C.F.J. Wu, "Resampling Inference with Complex Survey Data," *Journal of the American Statistical Association, 83* (1988), 231–241.

[26] J.F. Beaumont and Z. Patak, "On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling," *International Statistical Review*, *80*(1) (2012), 127–148.

In total, 500 bootstrap weights were created (*bsrwgt1, … bsrwgt500*), with the zeroth bootstrap weight (*bsrwgt0*) being the same as the students' final weight (*student_weight_final_part*), as many programs prefer this structure of bootstrap weights.

Note that for strata with a census of schools and a participation rate of 100 percent, $a_{h,k} = 1$. These strata are shown in Table A.6.

**TABLE A.6 School allocation by census strata**

| Stratum | Number of schools |
|---|---|
| British Columbia – francophone | 13 |
| Manitoba – francophone | 16 |
| New Brunswick – anglophone | 83 |
| New Brunswick – francophone | 60 |
| Newfoundland – francophone | 1 |
| Nova Scotia – anglophone | 114 |
| Nova Scotia – francophone | 11 |
| Prince Edward Island – anglophone | 22 |
| Prince Edward Island – francophone | 4 |
| Saskatchewan – francophone | 6 |
| "Volunteer" schools and additions | 3 |

The "volunteer" schools were added after sample selection — i.e., additions to the originally selected sample with something other than the pre-selected replacement schools. They form their own strata, as per Kish's[27] recommendation for "surprises."

All students have a response status. The possible status codes are provided in Table A.7 and were used to determine which students were respondents.

[27] L. Kish, *Survey Sampling* (John Wiley & Sons, 1963).

## TABLE A.7 Codes for participation status

| Participation code | Code description | Student response status |
|---|---|---|
| 1 | Absent | Non-respondent |
| 2 | Participated during scheduled session | Participant |
| 2A | Participated during scheduled session with an accommodation | Participant |
| 3 | Participated during makeup session | Participant |
| 4 | Exempted by school | Excluded |
| 5 | Exempted because appropriate modification could not be made | Excluded |
| 6 | No longer enrolled in this school/class | Left permanently |
| 7 | Parents and/or students who do not wish to write | Non-respondent |
| 8 | Not a Grade 8/Secondary II student | Excluded |
| 9 | Home-schooled student | Excluded |
| 10 | Answer sheet and booklet were not returned; only the questionnaire data were available | Non-respondent |
| 11 | Student responded to fewer than 3 achievement items per domain and did not complete at least the first section (Section 1) of the contextual questionnaire | Non-respondent |