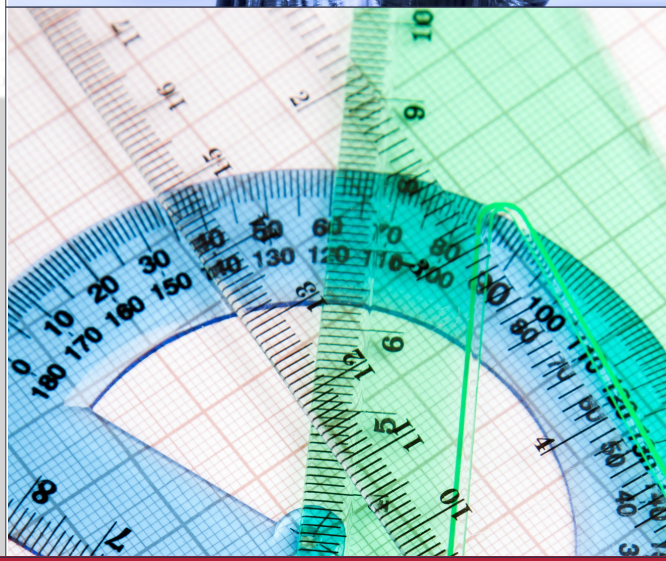
Pan-Canadian Assessment Program

# PCAP 2016

## Technical Report



cmec

Pan-Canadian Assessment Program

# PCAP 2016

## Technical Report

**Authors**

**Kathryn O'Grady**, Council of Ministers of Education, Canada

**Karen Fung**, Council of Ministers of Education, Canada

cmec

Council of Ministers of Education, Canada

Conseil des ministres de l'Éducation (Canada)

The Council of Ministers of Education, Canada (CMEC) was formed in 1967 by the 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, including education in early childhood and at the elementary, secondary, and postsecondary levels, and adult learning.

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.[1] 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 reading, mathematics, and science at the Grade 8/Secondary II[2] 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.*

[1] In this report, “ministry” includes “department.”
[2] PCAP is administered to students in Secondary II in Quebec and Grade 8 in the rest of Canada.

# Table of Contents

**Chapter 1. Pan-Canadian Assessment Program: An Overview ..... 7**
Context ..... 7
Pan-Canadian assessment ..... 7
Participation ..... 8
Administration time ..... 8
PCAP in both official languages ..... 8

**Chapter 2. Design and Development of the Assessment ..... 9**
PCAP assessment cycle ..... 9
Reporting PCAP achievement over time ..... 10
Updating the assessment framework ..... 10
Assessment design ..... 11
Working groups ..... 12
Updating the reading framework ..... 12
Item development ..... 12
Item translation and review ..... 13
Verification of the assessment items and coding guides ..... 13
Translating and comparing items in English and French ..... 14
Editing for language and style ..... 14
Technical editing ..... 14
Psychometric editing ..... 15
Advisory panel to examine test fairness ..... 16
GUIDELINES FOR ADVISORY PANEL TO EXAMINE TEST FAIRNESS ..... 17
Item approval by the provinces ..... 18

**Chapter 3. Development of the Contextual Questionnaires ..... 19**
Updating the questionnaire framework ..... 19
Gender identity ..... 20
Gender differences ..... 20
Confidentiality ..... 20

**Chapter 4. Sampling Procedures ..... 21**
Sampling design ..... 21
Sampling frames ..... 22
Stratification ..... 23
Sample sizes ..... 25
First stage sampling – sampling of schools ..... 26
Databases on the schools ..... 26
Selection of schools ..... 26
Exclusion of schools ..... 26
Second stage sampling – sampling of students ..... 27
Weighting of sample ..... 28

**Chapter 5. Field Testing ..... 29**
Item-review working group – field study ..... 29
Assessment booklets ..... 29
Scoring session ..... 29
Data capture ..... 29
Data analysis ..... 30
Item-selection working group – main study ..... 30
Review of the assessment framework ..... 31

**Chapter 6. Main Study ..... 32**
Assessment booklets ..... 32
Reviewing the assessment material ..... 32
Printing the assessment booklets ..... 32
Checking documents ..... 32
Letter sent to parents/guardians of students ..... 33
Administration procedures ..... 33
Assessment site ..... 33
Administering the assessment ..... 33
Students with special needs ..... 34
Questionnaires for the principal and teachers ..... 35
Participation and exemption of students from the assessment ..... 35
Organizing a makeup session ..... 37
Returning assessment materials ..... 37
Scoring session ..... 37
Bundling booklets ..... 37
Scoring sheets ..... 38
Scorers' manual ..... 38
Coding guide ..... 38
Scorer leaders ..... 38
Table leaders ..... 39
Scorer training ..... 39
Scoring reliability ..... 39
Reliability reviews ..... 39
Inter-rater reliability (double scoring) ..... 41
Trend reliability ..... 42
Reports and feedback ..... 42
Provincial Coordinator's Report ..... 42
School Coordinator's Report ..... 44
Scorer feedback forms ..... 46

**Chapter 7. Setting a Performance Standard ..... 47**
Standard-setting sessions ..... 47
Selection of an expert panel ..... 47
Preliminary performance-level descriptors ..... 47

Security of materials ..... 48
The Bookmark procedure ..... 48
Standard-setting procedure ..... 49
Performance-level descriptors ..... 51

**Chapter 8. Processing PCAP Data ..... 53**
Data gathering ..... 53
Data capture ..... 53
Data entry quality control ..... 53
Data cleaning ..... 54
General recoding ..... 54
Review of the sampling data ..... 54
Final review of the data and preparing the database ..... 54

**Chapter 9. Analysis of Achievement Data ..... 56**
Preliminary analysis ..... 56
Data screening ..... 56
Item recoding ..... 57
Missing data ..... 58
Not-administered items ..... 58
Not applicable items ..... 58
Not-reached items ..... 58
Omitted items ..... 59
Invalid response ..... 59
Item analysis ..... 59
Classical theory item analysis ..... 59
Item difficulty ..... 60
Item discrimination ..... 60
Specific statistics for MC items ..... 60
Specific statistics for CR items ..... 60
Examining for missing data ..... 61
Reliability of the PCAP 2016 assessment ..... 61
Problematic items ..... 61
IRT analysis ..... 62
Assessing the dimensionality of PCAP 2016 ..... 62
Item calibrations and assessing the IRT models' fit ..... 62
Differential item functioning ..... 63
Test functioning ..... 63
Linking and equating the minor domains with previous assessments ..... 63
Achievement score generation and scale scores ..... 64
Standard error estimates ..... 64
Presentation of the PCAP 2016 achievement results ..... 65

**Chapter 10. Analysis of Questionnaire Data ..... 66**
Preliminary analysis ..... 66

Data screening ..... 66
Item recoding ..... 67
Missing data ..... 67
Descriptive statistics ..... 67
Factor analysis ..... 68
Analyses of items and indices ..... 68
Group comparison analysis ..... 68
Correlational analysis ..... 68

**Chapter 11. PCAP Data Sets ..... 70**
Description of the data sets ..... 70
Student data set ..... 70
Teacher data set ..... 70
School data set ..... 71
Merged data set – student/teacher/school ..... 71
Accessing the data set for research ..... 71
Terms and conditions ..... 72
Information for researchers ..... 73

**References ..... 74**

## List of Tables

Table 2.1 Schedule of PCAP assessments ..... 9
Table 2.2 Number of clusters, scenarios, and items by domain and booklet ..... 11
Table 2.3 Distribution of items by type and assessment domain ..... 11
Table 2.4 Guidelines for advisory panel to examine test fairness ..... 17
Table 4.1 Coverage of the PCAP target population ..... 22
Table 4.2 The 2016 PCAP sample size allocation ..... 24
Table 4.3 Estimating the size of a sample ..... 25
Table 4.4 Summary of the criteria used to sample students, based on the types of strata ..... 28
Table 6.1 Reliability review results for reading scoring group 1 ..... 41
Table 6.2 Reliability review results for reading scoring group 2 ..... 41
Table 6.3 Reliability review results for mathematics ..... 41
Table 6.4 Reliability review results for science ..... 41
Table 6.5 Overal agreement between scorers for double scoring ..... 42
Table 7.1 Composition of the PCAP test booklets for reading items ..... 49
Table 7.2 Distribution of students by performance level in reading ..... 51

# Chapter 1. Pan-Canadian Assessment Program: An Overview

## 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, and 2018 Programme for International Student Assessment (PISA) (involving over 80 countries and economies in 2018), the 2011 and 2016 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 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 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 differ from one part of the country to another, 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 across the country.

PCAP is designed as a system-level assessment to be used primarily by provincial ministries of education to examine their educational systems. Information gathered by each assessment has given ministers of education a basis for examining curricula and other aspects of the school systems. PCAP data is 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 the assessment is 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 and 2016. Northwest Territories previously participated in SAIP.

**Administration time**

Students were allotted 90 minutes to respond to the PCAP 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 items in the assessment booklet, 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 the francophone school system of their respective provinces are reported as French. The results obtained from students educated in the anglophone school system of their respective provinces are reported as English. The majority of students in French-immersion programs wrote the assessment in English rather than assessing their reading literacy skills in their second language. As a resource for French-immersion students, a list of common science and mathematics terms was provided in English and French.

## Chapter 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 in the eight provinces that had sufficient numbers for valid statistical comparisons. Owing to the small sample size, results for students enrolled in francophone schools in Prince Edward Island were not indicated in the results; however, they were included in the calculations for the overall result in that province. Although the Saskatchewan francophone sample was also very small, it represented the majority of the Saskatchewan Grade 8 francophone population. Reporting of data for this population was approved by the Saskatchewan Ministry of Education.

Francophone students in Newfoundland and Labrador did not participate in PCAP 2016.

### 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, 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 fourth assessment, in 2016, the focus was on reading, as it had been in the first assessment, in 2007, with mathematics and science as the minor domains.

Table 2.1 Schedule of PCAP assessments

| Domain | Actual and Proposed Dates of PCAP Assessments | | | | | |
|---|---|---|---|---|---|---|
| | Cycle 1 | | | Cycle 2 | | |
| | Spring 2007 | Spring 2010 | Spring 2013 | Spring 2016 | Spring 2019 | Spring 2022 |
| Major | Reading | Mathematics | Science | Reading | Mathematics | Science |
| Minor | Mathematics | Science | Reading | Mathematics | Science | Reading |
| Minor | Science | Reading | Mathematics | Science | Reading | Mathematics |

**Reporting PCAP achievement over time**

One of the strengths of PCAP is its measurement of changes over time of 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 "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, and 2016, 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 2016 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 of students at a comparable point in their schooling. The *PCAP 2016 Assessment Framework* (CMEC, 2016)[3] provides the theory, design principles, and performance descriptors that were used to develop test items in each of the three domains for the second cycle of PCAP (2016–22). Chapter 2 describes the major domain of reading, while chapters 3 and 4 describe the minor domains of mathematics and science respectively.

For 2016, the reading and mathematics frameworks were updated to better reflect curricula and standards across Canada. 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 that cycle. The updates were discussed and revised by provincial experts.

A new content domain was added to the reading framework for the 2016 assessment. The world of digital information requires readers to bring skills and effort to their task in order to build coherent knowledge out of numerous pieces of media, including text and images. This ability to build coherence is a new challenge to literacy education beyond anything faced in earlier times, and here we call it "transliteracy." One context in PCAP 2016 involved introducing a complex task that involved multiple representations of information (i.e., text, diagram, graph,

[3] http://www.cmec.ca/537/Programs-and-Initiatives/Assessment/Pan-Canadian-Assessment-Program-(PCAP)/PCAP-2016/General-Information/index.html

map) and the assessment focused on the student producing a coherent account of the focal event, phenomenon, or problem.

**Assessment design**

The PCAP assessment, a paper-and-pencil test, covers three assessment domains: reading, mathematics, and science. Reading was the major domain, while mathematics and science were minor domains in the 2016 study. Just as with PISA, the focus changes with each assessment, so reading will become a minor domain and mathematics the major domain in the next PCAP study in 2019.

For the PCAP assessment, eight clusters of reading assessment units were distributed within four booklets, so that each booklet typically contained two clusters of reading items, one mathematics cluster, and one science cluster. One booklet contained a cluster of questions that were multidisciplinary and that were used to access students' knowledge and skills in reading, mathematics, and science using a single context. The four booklets were randomly and equally distributed to students within a single class. Thus, every student completed two of the eight clusters of reading 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. All the assessment booklets contained a student questionnaire at the end of the booklet.

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

Table 2.2 Number of clusters, scenarios, and items by domain and booklet

| | Reading | | | Mathematics | | | Science | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Clusters** | **Passages** | **Items** | **Clusters** | **Scenarios** | **Items** | **Clusters** | **Scenarios** | **Items** |
| **Booklet 1** | 2 | 4 | 25 | 1 | 4 | 8 | 1 | 3 | 9 |
| **Booklet 2** | 2 | 5 | 25 | 1 | 4 | 9 | 1 | 3 | 9 |
| **Booklet 3** | 2 | 6 | 26 | 1 | 4 | 9 | 1 | 3 | 7 |
| **Booklet 4** | 2 | 5* | 17 | 1 | 4* | 12 | 1 | 4* | 12 |

*includes one multidisciplinary scenario

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** |
| Reading | 19 | 6 | 20 | 5 | 18 | 8 | 15 | 3 |
| Mathematics | 4 | 4 | 5 | 4 | 5 | 4 | 7 | 5 |
| Science | 7 | 2 | 5 | 4 | 4 | 3 | 9 | 3 |
| Total | 30 | 12 | 30 | 13 | 27 | 15 | 31 | 11 |

### Working groups

The primary focus of PCAP 2016 was reading. Thus, new cognitive items were developed for that domain only, and the questionnaires were revised to focus on the teaching and learning of reading. Working groups consisted of experts in language arts 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 scoring sessions for the field test and main study.

### Updating the reading framework

Carl Bereiter and Marlene Scardamalia, from Knowledge Building Concepts, Incorporated, were commissioned to update the reading framework. This framework had been developed at the beginning of the PCAP program in 2007, based on current research related to literacy teaching. The revised framework reflects current thinking in the field of literacy research, including the assessment of reading.

The reading framework review working group reviewed and further revised the framework to reflect changes in the language arts program of studies in provinces across Canada. The working group, which met in Toronto in June 2014, comprised language arts curriculum and assessment specialists from Alberta, Quebec, New Brunswick, and Prince Edward Island.

### Item development

Documentation to guide all stages in the item-development process was prepared for the meeting of test developers which took place in Toronto in August 2014.

Provinces and territories were invited to nominate item developers and this working group had representatives from British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick, Nova Scotia, Prince Edward Island, and Newfoundland and Labrador.

The orientation included an overview of the reading 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 scoring 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 reading framework.

At the conclusion of this item-development session, the working group reviewed and revised the reading 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 reading items from previous administrations to ensure that they adequately represented the framework and were properly classified.

**Item translation and review**

The items were developed in both English and French and were translated and copy-edited at CMEC.

An item review working group meeting was held in Toronto in January 2015; the working group was made up of representatives from Ontario, New Brunswick, and Prince Edward Island. The meeting's goal was to review the reading items 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. The working group made one of three recommendations for each item: to a) keep the item unchanged, b) remove the item from the field test bank, or c) keep the item with minor changes as recommended by members of the working group. As a result of` this work, completed units consisted of the stimulus material and between two and six items. Units could contain one type of item (i.e., selected or constructed response) or a mix of both response types.

At the end of this working group meeting, 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 reading items based on the reading framework
- verifying the scoring guides for constructed response items

- editing for language and style, technical editing, and psychometric editing

### *Translating and comparing items in English and French*

Units of items, developed in both official languages by the item-development working group, were cross-translated by CMEC translators.

In a broad-scale assessment like PCAP, it is vital that the various versions of the test are parallel in terms of language to avoid giving one group an advantage over another. Although an assessment can always include differences between items, it was important to ensure that the items in the English version and those in the French version were as equivalent as possible. Additionally, any text assumes that students will have a degree of reading literacy. In PCAP, context or passage selections were chosen to be accessible to the vast majority of Grade 8/Secondary II students. Bilingual working groups of experienced educators reviewed and validated the items at each stage of development to ensure that the vocabulary was consistent with the level of understanding that can be expected of these Canadian students.

### *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; each item began with a statement followed by a question; the number of lines for the student's answer were 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 science, reading, and mathematics conducted a psychometric edit of items. For selected-response items, one factor to be checked was the order of the possible answers. In reading, possible answers could begin with shortest and end with the longest, thus 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 and answer the item correctly. 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 language arts. The coding guides for trend items remained unchanged from the previous administration in which each subject was the major domain (2007 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. For mathematics and science, codes could be 0 or 1, or 0, 1, and 2. In reading, 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, scorers 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.

**Advisory panel to examine test fairness**

At the 106th ACDME meeting in June 2015, deputy ministers approved the establishment of a panel to provide feedback on the appropriateness of the assessment materials used in national and international assessments facilitated by CMEC. The main purpose of the panel's work was to ensure that questionnaire items and passages or scenarios used for assessment items are free of bias and are as valid and fair as possible for a wide range of students.

The guiding principle for the panel was that nothing on an assessment should cause a student to feel so upset or distracted that they are unable to communicate their attitudes or understanding of the topic being explored. Assessment contexts and questionnaire items should reflect the values of inclusive public education in that they are respectful of diversity with respect to race, colour, religion, national origin, ancestry, place of origin, age, disability, marital status, real or perceived sexual orientation and/or gender identity, sex, social condition, or political belief or activity.

The advisory panel was asked to consider nine elements in their review of the context and questionnaire items as shown in Table 2.4.

Table 2.4 Guidelines for advisory panel to examine test fairness

GUIDELINES FOR ADVISORY PANEL TO EXAMINE TEST FAIRNESS

1. Universality
   - Are appropriate for all participating students, teachers, principals, and parents (including subpopulations)
   - Limits unfamiliar words, idioms, and idiomatic expressions
   - Avoids language or phrases that may be objectionable to a particular group
2. Gender
   - Does not include themes, subject matter, or organizational tone that favours one gender
   - Does not portray gender roles in a stereotypical or pejorative manner
3. Geography
   - Does not advantage or disadvantage students in certain parts of Canada
4. Socioeconomics
   - Avoids inappropriate emphasis on wealth, poverty, and crime
   - Does not suggest that affluence or poverty is attributed to a particular group
   - Does not suggest that belonging to a particular socioeconomic group is more advantageous than another
5. Social and environmental issues
   - Does not include references to issues such as violence, gambling, addictions, criminal behaviour, homelessness
   - Avoids topics such as family conflicts
   - Avoid focus on warfare, lives of soldiers (past or present)
6. Race, ethnicity, and culture
   - Does not portray ethnic groups in a pejorative or stereotypical manner
   - Does not include topics that suggest one culture is preferred or superior to another
   - Does not include cultural iconographic references to celebrations, national holidays (other than Canadian), or contests/prizes
7. Religion
   - Avoids topics that require a student to assume a position that may be contrary to their own religious beliefs or teaching
   - Avoids religious references
8. Disability
   - Includes people with disabilities in a natural and authentic manner
   - Does not portray disability or assumptions about a person's perceived challenges as a focus
9. Age
   - Does not portray an age group in a stereotypical or pejorative manner
   - Does not portray one age group as more favourable than another

Members of the Test Fairness Panel were asked to review the assessment materials for biases related to their specific area of expertise and provide guidance to the development team for refinement of the items. Members made recommendations with respect to which items should be accepted without change, accepted with changes, or rejected. These recommendations were used to prepare the final drafts of the assessment materials.

**Item approval by the provinces**

Before including items in the field test and main study, the provinces had to approve of the items selected. CMEC produced three field test booklets and four main study booklets, in English and French, and then sent these 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 2015 and the main study in 2016.

# Chapter 3. Development of the Contextual Questionnaires

Students participating in PCAP, in addition to their teachers and school principals, complete questionnaires which are designed to provide provinces and territories with contextual information that would contribute to the interpretation of the performance results. The information from these questionnaires may be used by researchers, policy-makers, and practitioners to help determine what factors influence learning outcomes. The content of the contextual questionnaires changes depending on which of the three domains is the primary focus in a PCAP assessment.

Contextual questions accompanying the PCAP 2016 assessment include factors that have been found in past studies to correlate with student achievement. Some examples of these correlates include:

- parental level of education
- language spoken in the home
- number of books in the home

The contextual questionnaires also communicate teaching and learning conditions, including teachers' homework expectations, areas of specialization, and years of teaching experience.

More details about the conceptual framework of the questionnaire component of PCAP are described in chapter 5 of the *PCAP 2016 Assessment Framework*. The conceptual framework reflects current research findings and best practices in the field of literacy development and the learning of reading. It also includes information gathered from questionnaires (student, teacher, and school) to capture contextual data.

**Updating the questionnaire framework**

Reading experts were asked to independently review and revise the questionnaire framework used for PCAP 2007, the previous year in which reading was the focus of the assessment. The experts were asked to identify important aspects of the context in which students learn to read and become literate, as well as to identify areas that could be improved or topics that could be added based on the academic literature on literacy, which contextualizes how students learn to read 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 reading, i.e., the major domain, and that could provide important contextual information for the provinces and territories. The reading-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 2016 assessment: one for participating students, one for their Grade 8/Secondary II language arts teachers, and one for school

principals. These questionnaires also focused on the particular need to capture factors associated with reading 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, reading, 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 reading.

**Gender identity**

Inclusive education is valued in Canadian provinces and territories and has led to development of policies and resources to support inclusion. One aspect of inclusive education is gender identify. In the PCAP 2016 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?

- Male
- Female
- I identify myself in another way.
- I prefer not to say.

**Gender differences**

Differences in reading achievement favouring girls 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;
- differential reading-related behaviours or interests outside of school.

Although this issue is less strongly emphasized for mathematics and science, there remains an interest in following trends in gender differences over time.

**Confidentiality**

Both the teacher and school questionnaires were linked to student results but used unique identifiers to preserve confidentiality. .

# Chapter 4. Sampling Procedures

In the spring of 2016, the fourth Pan-Canadian Assessment Program was administered. It assessed three domains: reading, mathematics, and science with reading being the primary domain. Four assessment booklets were used in which all three domains were assessed, with the majority of the items focusing on reading. One school grade—Grade 8/Secondary II—was assessed. Eighteen populations 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 and 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 2016, 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 2016; greater detail on sampling and weighting can be found in the report prepared by Statistics Canada (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[4] facilitates the sampling process and prevents ambiguities. Table 4.1 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. 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 2015 field test of the assessment.

[4] "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.

Table 4.1 Coverage of the PCAP target population

| | Population Projections* (13-year-olds) | | Frame** | | Proportion of Population (%) | |
|---|---|---|---|---|---|---|
| **PROVINCE** | **2013** | **2015** | **2013** | **2015** | **2013** | **2015** |
| British Columbia | 47,129 | 45,848 | 46,895 | 45,576 | 99.5 | 99.4 |
| Alberta | 45,386 | 45,354 | 40,094 | 46,578 | 88.3 | 102.7 |
| Saskatchewan | 13,677 | 13,371 | 12,598 | 12,335 | 92.1 | 92.3 |
| Manitoba | 16,289 | 15,852 | 14,451 | 14,174 | 88.7 | 89.4 |
| Ontario | 153,091 | 147,322 | 145,756 | 146,765 | 95.2 | 99.6 |
| Quebec | 79,921 | 78,354 | 85,435 | 81,563 | 106.9 | 104.1 |
| New Brunswick | 8,010 | 7,543 | 7,976 | 7,327 | 99.6 | 97.1 |
| Nova Scotia | 9,785 | 9,146 | 9,792 | 8,674 | 100.1 | 94.8 |
| Prince Edward Island | 1,657 | 1,527 | 1,487 | 1,433 | 89.7 | 93.8 |
| Newfoundland and Labrador | 5,436 | 5,221 | 5,441 | 5,193 | 100.1 | 99.5 |
| **CANADA** | **380,381** | **369,538** | **369,925** | **369,618** | **97.3** | **100.0** |

*Source: Statistics Canada. Table 051-0001 – 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.

## 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 2016. 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.[5] 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 bases on the following a priori categories:

- special schools in which all students had special education needs,

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

- 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.

**Stratification**

*Stratification* is a means of organizing the sampling frame so that a 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 at 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 province by language of 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 school 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 4.2 outlines the list of strata and whether schools were sampled at stage one, or whether censuses were used at the schools and/or student level.

Table 4.2 The 2016 PCAP sample size allocation

| Stratum | Sampling at the First Stage | Census of Schools | Census of Students | Expected Number of Schools | Expected Number of Students |
|---|---|---|---|---|---|
| British Columbia – anglophone | X | | | 150 | 3,011 |
| British Columbia – francophone | | X | | 12 | 171 |
| Alberta – anglophone | X | | | 150 | 2,990 |
| Alberta – francophone | | X | | 20 | 299 |
| Saskatchewan – anglophone | X | | | 186 | 3,143[6] |
| Saskatchewan – francophone | | X | X | 5 | 79 |
| Manitoba – anglophone | X | | | 150 | 2,846 |
| Manitoba – francophone | | X | | 19 | 267 |
| Ontario – anglophone | X | | | 150 | 2,713 |
| Ontario – francophone | X | | | 125 | 2,094 |
| Quebec – anglophone | | X | | 89 | 1,801 |
| Quebec – francophone | X | | | 150 | 3,361 |
| New Brunswick – anglophone | | X | | 83 | 1,516 |
| New Brunswick – francophone | | X | | 61 | 1,164 |
| Nova Scotia – anglophone | | X | | 116 | 2,480 |
| Nova Scotia – francophone | | X | | 11 | 184 |
| Prince Edward Island – anglophone | | X | | 22 | 438 |
| Prince Edward Island – francophone | | X | X | 3 | 58 |

[6] Confusion regarding the categorization of dual schools in Saskatchewan with the anglophone or francophone school districts resulted in a larger sample of anglophone students than required by the sampling framework. Dual schools in Saskatchewan are under the mandate of the anglophone school system.

| | | | | | |
|---|---|---|---|---|---|
| Newfoundland and Labrador – anglophone | | X | | 114 | 1,865 |
| Newfoundland and Labrador – francophone | | X | X | 3 | 23 |

*Note:* Although students in francophone schools in Prince Edward Island were sampled, the results were not reported due to small sample size. Although students in francophone schools in Newfoundland and Labrador were considered in the sample allocations, these students did not participate in PCAP 2016.

**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 per cent overall, with a confidence level of 95 per cent, which was consistent with previous PCAP administrations. Table 4.3 gives the formula used to determine the size of a sample in relation to the calculation of frequency distributions.

Table 4.3 Estimating the size of a sample

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

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 the 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 per cent) 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 per cent 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 2016. The same lists prepared for step 5 were used, with reasons given for any student's non-participation in the assessment.

Table 4.4 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><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.

# Chapter 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 working group – field study

The item-review working group, with representatives from four provinces, met to review all reading items for content, vocabulary, and translation, and to verify the items' classification for the test's subdomains. This group also reviewed the items to identify any issues with the vocabulary level and for biases (e.g., gender, culture, geography). The working group selected items for the field test that represented a broad coverage of the reading subdomains and with a range of difficulties.

### Assessment booklets

Three booklets were compiled for the field test. Each booklet followed the specifications outlined in the *PCAP 2016 Assessment Framework* and contained about 40 reading items in addition to the Student Questionnaire. The students had 90 minutes to complete the booklet and 30 minutes for the Student Questionnaire. Teacher and school questionnaires were prepared as separate booklets.

### Scoring session

The scoring session took place over five days in Ottawa in July 2015. 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.

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

### Data capture

Students recorded their selected-response answers (e.g., multiple choice, true or false) on a tear-out answer sheet and wrote their answers for constructed response questions directly in the PCAP 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.

Scorers at the field-test scoring session coded only the constructed-response items by filling in the circle on a coding sheet that best matched the student's response. Booklets were distributed to scorers in bundles of 10, with booklets from various provinces in each bundle. Following the scoring session, all booklets were shipped to a data-capture firm that captured data from selected-response items and from Student, School, and Teacher Questionnaires, as well as scorer codes for constructed-response items. Collected data were merged in an Excel file to create a database.

**Data analysis**

Field-test data for the reading items were analyzed simultaneously by a CMEC expert on psychometrics and an external expert. The parallel process was to ensure validation of the data. 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).

One field-test database was created for assessment and questionnaire items. In addition to an overall achievement score, scores for each scenario and subdomain were produced.

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 gender and language.

**Item-selection working group – main study**

Following field testing, the item-selection working group, with representatives from seven 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-scoring session (from the questionnaire administered at the end of the scoring 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: reading, mathematics, and science. All reading, mathematics, 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 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 working group also 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 2016 Assessment Framework* was reviewed to ensure alignment between the framework and assessment items. Very few changes to the framework were required.

# Chapter 6. Main Study

The PCAP assessment was conducted between April 25 and May 20, 2016,[7] with the primary domain being reading. (The minor domains were mathematics and science.) More than 30,000 students selected at random from over 1,500 Canadian schools in 10 provinces took part in the test in English and French. The main items assessed the knowledge and skills of Grade 8/Secondary II students in all three subject areas.

### Assessment booklets

Each assessment booklet included two clusters of reading items, and one cluster each of mathematics 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.

### Printing the assessment booklets

After all provinces approved of them, sample assessment booklets were printed for review to ensure that all changes made to the content by the working groups, as well as by CMEC, had been incorporated into the new versions. 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 them either to the provincial coordinators for distribution to the schools or directly to the school coordinator.

### Checking documents

Each school or provincial coordinator had to ensure that he or she had the materials for each school. Any missing packages 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.

[7] Administration dates differ in Saskatchewan, Quebec, and Prince Edward Island.

**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) that was printed on the cover of the assessment booklet and on the tear-out answer sheet and scoring sheet. The IDs were assigned to protect students' confidentiality. Students' names from the *List of Students* were used to facilitate the booklets' administration process in schools. If a student joined the selected class after the sampling was performed, they were allowed to participate in the evaluation by using extra booklets that were provided.

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, although the test cannot be made available in an electronic format for visually impaired students, large-print test formats could be produced to accommodate their needs.

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 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 to each student assigned 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 has been absent from the assessment.

Once the test was completed, the coordinators had to collect all assessment documents and store them in a secure place to keep the material confidential.

The provincial coordinators also had to ensure the assessment's smooth administration. They were responsible for observing the assessment's administration in between 5 and 10 per cent 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. 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 these special test versions were included in the shipment of assessment booklets to the school. 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, large print, coloured paper, 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 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**

The School Questionnaire was usually filled out by the school principal. The selected classes' science teachers had to fill out the Teacher Questionnaire. Both questionnaires were available in both paper and on-line versions. 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 language arts teachers associated with the selected students were asked to fill out a questionnaire (one questionnaire per teacher). Each questionnaire had an identification number written on the cover page. A Teacher Questionnaire ID number was assigned to each teacher identified and printed on the front cover of the questionnaire. The questionnaires had to be distributed at the time of the assessment.

All the questionnaires were collected (or questionnaire covers for those completed on-line) at the end of the assessment session. Under no circumstances could the school coordinators reveal the teachers' names. They had to destroy any list of teacher names after the assessment to ensure confidentiality.

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. 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 reading, mathematics, 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 nine 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 = Students who do not wish to write
8 = No data available

The three exemption codes are given here.

**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:

- 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: must write what student says without editing)
- verbatim reading of instructions only, for all domains
- verbatim reading of occasional prompts and/or questions for science and mathematics only [in cases where the entire science and/or mathematics 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: Sampling Procedures and Response Rate, page 157 of the PCAP 2016 public report (O'Grady, Fung, Servage, & Khan, 2018).[8]

**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 A (absent) and B (participated during scheduled session) codes and calculate the percentage rate for student participation using the following formula:

$$\frac{(B)}{(A+B)} x\ 100$$

If the student participation rate was less than 85 per cent, a makeup session had to be held before whatever dates were indicated. 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, his or her participation code changed from A (absent) to C (participated during makeup session) on the Student Tracking Form.

**Returning assessment materials**

After assessing the students, the school coordinators had to fill out the School Coordinator's Report. They also had to fill out the School Packing List and indicate the number of each 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, all the School Questionnaires, the Teacher Questionnaires, the assessment booklets and answer sheets, as well as the copies and photocopies of unused assessment booklets.

**Scoring session**

The scoring session for the main administration was held in Ottawa, Ontario, in July 2016. All the reading, mathematics, and science items were scored by teachers in the relevant domains. Reading-item scoring was scheduled for two weeks because this was the primary domain and there were more items to correct. The scoring of mathematics items was scheduled for the first week only, while science items were scored during the second week. In all, there were 108 scorers, both anglophone and francophone.

In total, approximately 27,000 assessment booklets were scored, with approximately 20,000 in English and 7,000 in French.

**Bundling booklets**

[8] https://www.cmec.ca/Publications/Lists/Publications/Attachments/381/PCAP-2016-Public-Report-EN.pdf

All the assessment booklets were bundled before scoring the items. A bundle contained 10 assessment booklets from various provinces. During the scoring sessions, the scorers picked up a bundle and verified the student ID code for each booklet. They could not obtain any information about the students' identity from the assessment books. (Because students were identified using a student ID code only, the assessment preserved their anonymity.) The scorers were also unable to determine which province the assessment booklets came from, to avoid any bias in scoring the items toward one province over another.

**Scoring sheets**

Tear-out scoring sheets for the constructed-response questions were located at the back of each assessment booklet. All constructed-response items were coded by scorers 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 scorers assigned various codes to the students' responses and recorded them on the scoring sheets. Once the scoring session was finished, the scoring sheets were returned to CMEC where the data were scanned and a database was created that contained all the assessment and questionnaire data.

**Scorers' manual**

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

**Scorer leaders**

The scorer leaders met for a few days in June to prepare for the scoring session. They reviewed and adapted the materials related to the assessment, such as the Coding Guide. They also prepared the training materials for the scorers. While preparing the training material, scorer leaders selected samples of student work to be used as examples or in training papers. These samples were scanned and inserted into the appropriate training document. 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. Scorer leaders were responsible for training table leaders and ensuring the smooth progress of the scoring session.

### Table leaders

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

### Scorer training

All scorers, including table leaders, received training on the coding guides for reading, mathematics, or science depending on their assigned scorer role, before scoring student papers. Prior to the training session, scorer 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 papers were then used to practise scoring and to further internalize the coding scheme. Pairs of scorers then scored a bundle of booklets and discussed the codes they assigned. This process was repeated for several bundles until their scoring was consistent with the coding guides. At the end of training, when scorers were able to consistently apply the coding standards, they proceeded with individual scoring until all assessment booklets were scored. Once the scoring of a scenario was completed, the scorers received training on the scoring of items in the next scenario.

To ensure high consistency in scoring, one question was coded in all 10 booklets before the scorers moved onto the next question in the cluster or scenario. This process was repeated until all questions in the cluster were coded. When the scorers had finished correcting a bundle, they returned it to the table leader and coded another bundle until the entire cluster, which could contain one or more scenarios, was completed. Throughout the scoring processes, table leaders did a random check of the codes assigned by each scorer to ensure consistent adherence to the coding guides. Issues that arose with specific questions were addressed by either individual or group retraining, and in a few cases, by recoding the question.

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

### Scoring reliability

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

#### *Reliability reviews*

In a scoring session, it is always important to implement the necessary procedures to ensure that scorers are coding correctly, because they must all agree on the various codes to ensure the results' validity. Prior to the scoring session, CMEC staff selected items at random from all the assessment booklets they received from provinces to conduct reliability reviews. The items

selected from one or more scenarios were then distributed to the scorer leaders for coding. Their responses were then returned to CMEC staff for entry into an Excel file used for comparison with scorer responses during the scoring session. If scorer 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 scorers. Reliability reviews were run for all anglophone or francophone scorers in all three domains. The reviews' goal was to monitor consistency throughout the scoring session. The reliability reviews occurred approximately twice per day and followed this procedure:

- At a time determined by the scorer leader, everyone stopped coding and coded the same student samples.
- Codes from scorers were compared to the benchmark (provided by the scorer leaders).
- Data were entered immediately by CMEC staff who provided results to the scorer leader.
- Scorer leaders debriefed the entire group or individual scorers.
- If the consistency was below 80 per cent on a specific question, individuals or groups of scorers were retrained and the booklets were rescored as required.

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

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

$$Percentage\ agreeement\ = \frac{Total\ number\ of\ agreed\ responses}{Total\ number\ of\ reliabilty\ tests} \times 100.$$

At the end of the scoring session, all the percentages obtained for each reliability review for each scorer were compiled. This constituted the total level of agreement as a percentage. The reliability tests for reading were divided into two groups, since it was the major domain, and therefore more of these items were required to be scored. Results showed that most scorers obtained a more-than-acceptable level of agreement with the experts. The reliability review results were satisfactory for all groups in mathematics and science. One of the two groups in reading obtained an overall percentage agreement below 80 per cent. The percentage agreement by language and overall for each scoring group and domain are presented in the tables 6.1 to 6.4 below.

Table 6.1 Reliability review results for reading scoring group 1

| Reliability Test | Cluster 1 | | | | | | | Cluster 2 | | | | | | % Agreement |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | |
| English | 60 | 76 | 73 | 77 | 71 | 88 | 87 | 56 | 77 | 72 | 80 | 77 | 77 | 76 |
| French | 74 | 79 | 88 | 69 | 86 | 84 | - | 80 | 78 | 88 | 69 | 81 | 90 | 71 |
| Overall | 67 | 78 | 80 | 73 | 78 | 86 | 87 | 68 | 78 | 80 | 74 | 79 | 84 | 74 |

Table 6.2 Reliability review results for reading scoring group 2

| Reliability Test | Cluster 1 | | | Cluster 2 | | | Cluster 3 | | | | % Agreement |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 4 | |
| English | 89 | 84 | 86 | 94 | 94 | 94 | 97 | 97 | 95 | 89 | 92 |
| French | 81 | 80 | 81 | 97 | 97 | - | 94 | 96 | - | - | 89 |
| Overall | 85 | 82 | 83 | 96 | 96 | 94 | 96 | 97 | 95 | 89 | 91 |

Table 6.3 Reliability review results for mathematics

| Reliability Test | Cluster 1 | Cluster 2 | | | Cluster 3 | | Cluster 4 | | % agreement |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 1 | 2 | 3 | 1 | 2 | 1 | 2 | |
| English | 99 | 97 | 94 | 98 | 91 | 97 | 90 | 99 | 96 |
| French | 95 | 97 | 95 | 94 | 100 | 90 | 99 | 97 | 96 |
| Overall | 97 | 97 | 95 | 96 | 96 | 93 | 94 | 98 | 96 |

Table 6.4 Reliability review results for science

| Reliability Test | Cluster 1 | | | Cluster 2 | | Cluster 3 | | Cluster 4 | | % Agreement |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 1 | 2 | 1 | 2 | |
| English | 83 | 95 | 90 | 88 | 83 | 66 | 72 | 82 | 79 | 81 |
| French | 84 | 94 | 81 | 88 | 81 | 69 | 69 | 79 | 73 | 80 |
| Overall | 84 | 94 | 85 | 88 | 82 | 68 | 70 | 81 | 76 | 81 |

***Inter-rater reliability (double scoring)***

Double scoring was a quality control measure in scoring assessment booklets for reading, mathematics, and science, in which about 2,870 booklets (approximately half of booklets in English and in French) were scored a second time by another scorer. Table 6.5 presents the overall consistency between scorers for the three domains. For reading, the agreement for Group 1 was lower than expected (66 per cent), while the overall agreement for Group 2 was acceptable (86 per cent). The lower-than-expected reliability and inter-rater reliability for

reading Group 1 will result in the removal of a trend reading context from future administrations of PCAP.

Table 6.5 Overall agreement between scorers for double scoring

| Booklet | Domain and Scoring Group | | | |
|---|---|---|---|---|
| | Reading Group 1 | Reading Group 2 | Mathematics | Science |
| **1** | 64 | 81 | 96 | 90 |
| **2** | 65 | 88 | 92 | 86 |
| **3** | 63 | 90 | 94 | 86 |
| **4** | 72 | * | 96 | 83 |
| **Average** | 66 | 86 | 95 | 86 |

** Due to fewer numbers of scored items in reading for Booklet 4, all items were grouped into Reading Group 1 for this analysis.*

### *Trend reliability*

Trend reliability was a quality control measure to estimate the degree of agreement between reading, mathematics, and science scorers for the anchor items in PCAP 2016 and the same items in PCAP 2013. Four items for reading, eight items for mathematics, and six items for science were common between the two administrations. About 1,910 booklets from 2013 (almost half in each language) were scored a second time by another scorer. When the items occurred in more than one booklet, then only one booklet was scored for trend reliability. Booklets were scored as determined by the scorer leaders. Trend-reliability scoring was done at the same time as the main scoring procedure but early in the process so that the data could be used to align scoring between the two administrations if required.

## 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. Scorers provided feedback during the scoring session. The information collected in these reports is summarized 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. Although provinces agreed to administer PCAP in English to French-immersion students, so that students wrote the reading assessment in their first language, some schools chose to administer the assessment in French to these students. The reason for this choice was that some schools indicated that such students writing in English were unfamiliar with the English vocabulary in the mathematics and science portions of the test, since their classroom instruction in these subjects had been in French.

Provincial coordinators also had to summarize the problems encountered by the schools during the assessment's administration. A small percentage of booklets had a page either missing or duplicated or staples improperly placed. There was some confusion for students when questions asked them to choose yes or no and then to justify their choice because some students thought they could *either* make their choice *or* justify it. Teachers indicated that they helped their students to understand this. Some schools received either booklets or questionnaires in the wrong language. Provincial coordinators were able to address some of these issues as they arose because they had extra booklets in both languages.

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, or because the number of test materials was incorrect. Some schools also expressed concerns about specificity and clarity of the information related to the administration of the test. There were schools that were not fully equipped to deal with a two- to three-hour testing session, and that would have required more support during 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 on one or two pages 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.

The provincial coordinators also suggested changes to the assessment for students with special needs or students who had problems with the language of the assessment. According to the reports, only modifications that are normally provided to special-needs students were made, including: provision of scribes, educational assistants (EAs) available to help with reading, an alternate location, extra time, large print, and English/French dictionaries for non-native speakers.

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.

***School Coordinator's Report***

After administering the assessment, the school coordinators had to fill out a report on how smoothly the assessment was administered. The School Coordinator's Report had 13 questions. The coordinators' comments will assist in better planning of future assessments while gathering information on how the administration of PCAP proceeded.

First, the school coordinators had to describe the measures taken to encourage students to participate seriously in the assessment. The measures used to encourage students to do their best were similar to those described earlier by the provincial coordinators. Many school coordinators reported that they took time with the selected students to talk about the assessment and explain the importance of doing their best on the test and responding to the items seriously. Some even met with students a few days before administering the assessment and went over the sample questions with them. Some school coordinators explained to students that the data obtained would support comparisons of results between provinces and that it was important to effectively represent their school and province. Others sent information about the PCAP assessment to parents or guardians of the students selected for the assessment so they would know the purpose of the assessment and could personally motivate their child. Some school coordinators said that some students were rewarded at the end of the assessment with a free snack such as pizza, which also appeared to motivate them to perform well on the test.

The school coordinators had to specify in their report whether they made any changes to the terms of the assessment for students with special needs. Some school coordinators had to give these students extra time to complete the assessment. Other students were placed in another classroom or a quieter place so they could concentrate better. In addition, some students were provided with readers (to read the questions verbatim to the students) or scribes.

The school coordinators also had to state whether there had been any problems during the assessment. The vast majority reported that the assessment session had run smoothly, though for some, various problems had arisen during the administration. Before administering the assessment, the school coordinators received the list of students on the Student Tracking Form with an identification number for each. Some invigilators did not read instructions carefully and handed out booklets randomly, only to find out later that there were assigned booklet numbers for students. Therefore, there was confusion over who had which booklet. They noticed that some students were tired and not motivated to write the test, because it took place at the end of the year (over 80 per cent of the tests were written in May) and they had to write a number of final exams. Thus, some school coordinators feared that certain students did not take the test seriously. Other school coordinators did not know that the students could have access to resource materials (e.g., dictionary, bilingual dictionary, manipulatives, thesaurus, or calculator), and they recommended that this be specified more clearly before the test's administration.

The school coordinators were asked to indicate the assessment procedures they were unable to follow. Most followed the assessment procedures, since no problems arose during the administration. Some, however, were unable to follow the procedures, as identified in the list below:

- Some schools chose not to notify students and parents of the test prior to the test date.
- Some schools inadvertently overlooked the instruction regarding the ID coding numbers of test booklets (matching specific ID codes to specific individual students). Where problems were discovered, the errors were corrected.
- Some invigilators lost the Student Tracking Forms and handed out booklets randomly. Therefore, there was confusion over who had which booklet.
- Some schools did not complete the Student Tracking Forms correctly, which led to later problems.
- Because several schools felt the time was too long for some students to focus or because of special events happening at the school, the assessment was broken up into two settings instead of being administered in one 90-minute setting with a short break.

The school coordinators were also asked to comment on the introductory scenarios (e.g., appropriateness, level of difficulty, interest level) and test questions (e.g., poor wording, more than one or no correct answer, age inappropriateness). Most school coordinators reported that the level of test items was appropriate and students were engaged. However, several felt that the questionnaires were too long and sometimes repetitive. A certain number of comments were also made about the wording of the test items.

The school coordinators also had to calculate the participation rate. If the student participation rate was less than 85 per cent, a makeup session had to be organized. According to the School Coordinator's Reports, the participation rate was above 85 per cent in almost all the schools.

### *Scorer feedback forms*

Following the scoring session for items from the main administration, approximately 120 scorers filled out a questionnaire that gathered their opinions and comments to help plan future scoring sessions and assessments. The scorer feedback form was divided into three sections. The first section contained the scorer's personal information, the second focused on the scoring process, while the third covered the assessment instrument.

Scorer feedback was generally positive regarding the material provided, the venue, the scoring process, and the assessment materials.

Scorers were asked to review the assessment questions to share their insights from the scoring session regarding students' strengths and weaknesses in the three domains. They were also asked to compile a list of common misconceptions presented by the students in their responses. The information will be included in a forthcoming issue of *Assessment Matters!* in which two reading passages from the PCAP 2016 Reading Assessment were released with commentary on the student work and sample responses.

# Chapter 7. 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 2016, reading was the major domain in PCAP for the second time, and significant changes were made to the assessment framework so it was necessary to establish performance standards.

### Standard-setting sessions

Standard-setting sessions took place in February 2017 in Toronto. The meetings were divided into three sessions: a one-day leaders' training session; two-day standard-setting sessions; and a one-day writing session to revise proficiency level.

The standard setting aimed to articulate levels of performance on the PCAP reading 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.

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 having some expertise in measurement and evaluation and in reading content for Grade 8/Secondary II. The standard-setting committee consisted of 21 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 21 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 2007, when reading was the major domain for the first time, as well as literature from international tests (e.g., TIMSS, PISA) were reviewed. These were statements describing what students at the three 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. The PLDs were stated in terms of the PCAP reading framework and at this first step, panellists made suggestions for revisions in consultation with each other and with the guidance of the CMEC facilitator. During the meeting the facilitator wrote the edits and suggestions and projected these on a screen so that they could be easily viewed. The facilitator then integrated the suggested language into revised PLDs that the panellists agreed upon. Once finalized, the PLDs were ready for use at the standard-setting meeting.

**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 2016 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 reading 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. See Table 7.1 for a breakdown of item format.

Table 7.1 Composition of the PCAP test booklets for reading items

| | Selected Response | Short Constructed Response (codes 0 or 1) | Extended Constructed Response (codes 0, 1, or 2; or codes 0, 1, 2, or 3) |
|---|---|---|---|
| Booklet 1 | 24 | 1 | 5 |
| Booklet 2 | 23 | 3 | 2 |
| Booklet 3 | 18 | 3 | 5 |
| Booklet 4 | 15 | 1 | 2 |

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, 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 OIB, panellists were asked to identify the item where students at each proficiency level will have a two-thirds chance of getting the item correct, and mark that item on the OIB with a bookmark—hence the 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 and between levels 2 and 3.

**Standard-setting procedure**

Twenty-one participants from all provinces and three CMEC staff members took part in the standard-setting session. Participants were assigned to two anglophone or two francophone tables. Each table had a leader and four or five 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 scoring 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 three 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 first bookmarking round took place before the end of the day, with participants reviewing each item in the OIB independently, then discussing as a group their conclusions and the reasons that one item was more difficult than ones ranked lower in the booklet. Following discussion, each participant would select 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 place a bookmark in the OIB. CMEC compiled all participants' responses by recording in an Excel file the OIB page numbers denoting the two cut scores. The mean of all responses defined the location of cut scores between levels 1 and 2 and between levels 2 and 3.

The second day began with a full group discussion on the first bookmarking round. CMEC staff posted the results with a graph that illustrated all the OIB page numbers identified by the panellists at the two 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, especially for the cut score between levels 1 and 2, with some placing the first cut-off at the very beginning of the OIB and others placing it much further into the booklet. These variations led to important and relevant discussions, with panel members explaining to each other why they had bookmarked a specific item. Several participants reported difficulty placing the first bookmark because they felt that some items were easier for students, while the data showed the opposite. Such items were therefore ranked further into the booklet. For the second cut score between levels 2 and 3, there was a smaller range of page numbers, indicating a greater degree of agreement. The first round was a good exercise for the panellists and gave them an opportunity to share comments and opinions.

The second round was similar to the first, with panellists placing bookmarks independently in the OIB to determine the two cut scores and providing a rationale for their choices as a group. The goal of the second round is to obtain a more coherent set of results than had been gathered in the first round. CMEC staff compiled results and shared them 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 both cut scores than in the first round.

Although the range of the lower cut score was smaller, improvement was still required to increase coherence. For the second round (but not the first), participants were shown impact data, that is, the percentage of students performing at levels 1, 2, and 3. Based on panel responses in the second round, approximately 45 per cent of students performed at level 1, 48 per cent at level 2, and 7 per cent at level 3. 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.[9] 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 7.2.

Table 7.2 Distribution of students by performance level in reading

| Performance Level | **Level 1** | **Level 2** | **Level 3** |
|---|---|---|---|
| Mean score in reading | 399 or below | 400 to 602 | 603 or above |
| Percentage of students | 14 | 73 | 12 |

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 texts and questions were appropriate and that the reading assessment was fair to Grade 8/Secondary II students.

**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 three levels of performance. Using these items, they developed a description of the knowledge and skills that

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

characterized achievement at each of the three performance levels.[10] Level 2 is considered the acceptable or "baseline proficiency," or the level at which students begin to demonstrate the level of reading literacy needed to participate in life situations. 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 3), and this progression in task difficulty/complexity applies both to overall reading and to each subdomain in the assessment.

[10] These descriptions appear in the PCAP 2016 public report (O'Grady, Fung, Servage & Khan, 2018) available at https://www.cmec.ca/publications/lists/publications/attachments/381/pcap-2016-public-report-en.pdf

# Chapter 8. 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.

### Data gathering

Assessment booklets and questionnaires were handed out during the test's main administration. Before data cleaning, in Canada, 27,484 Grade 8/Secondary II students wrote the assessment and responded to the Student Questionnaire; 1,470 English language arts teachers of the participating students completed the Teacher Questionnaire; and 1,355 school principals responded to the School Questionnaire. Data from these documents were gathered over a period of several weeks.

### Data capture

As in the field test, in the main study students filled in bubbles on an answer sheet for selected-response items or wrote out their answers in a few sentences in the assessment booklet for constructed-response items. Once they completed the assessment, students had 30 minutes to answer the Student Questionnaire at the end of the assessment booklet.

After administration of the main study was completed, the provinces sent all the assessment booklets, answer sheets, and questionnaires to CMEC for data capture. The Student Questionnaire, School Questionnaire, and Teacher Questionnaire contained selected-response items and did not have to be coded by experts, so they were sent to an external company for data capture. Assessment booklets were then shipped to Ottawa, Ontario, for the scoring session, where the scorers corrected all constructed-response items in more than 27,000 booklets. A code was assigned to each item by filling in the appropriate bubbles on a scoring sheet.

Two techniques were used for data capture. Data on bubble answer sheets were captured using an optical scanner. Manual data entry was used to capture questionnaire data.

For achievement data, files with unreadable data or items with multiple responses were identified by optical mark recognition software. For example, if the bubbles on a particular answer sheet were not darkened sufficiently, then the program identified this as a problem file. The data officer checked these electronic files individually and input the data manually.

### Data entry quality control

For questionnaire data, all the data were double entered by the data entry company. Any discrepancies were taken up to a third person to address the differences.

## Data cleaning

When data were received after the scoring 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.
- Duplicate data records were identified and only one record was kept, if the two records showed a 100 per cent 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. This 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 printing errors or translation errors, 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. Students, teachers, 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.

# Chapter 9. Analysis of Achievement Data

This chapter outlines the PCAP 2016 analysis of achievement data. It describes and identifies 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 2016 reading and mathematics with PCAP 2007, 2010, and 2013
5. scoring and scaling PCAP 2016 performance data
6. standard error estimates
7. presenting the PCAP 2016 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.[11]

In addition, data were cleaned so that cases with all blank responses were removed. Due to the design of the PCAP booklets, students who attempted at least one MC and one CR item per subject were kept for analyses to ensure that data were retained for those who made an effort to write the assessment. During this stage of cleaning, cases were identified where students wrote the assessment who were exempted (participation codes 4, 5, or 6)—either exempted by the schools, exempted because appropriate modifications could not be made, or the student 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 one item per question type per subject.

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

Cases of misidentification could have been the result of the student being given a booklet with an ID number that did not correspond to the Student Tracking Form. 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. During this process, it was found that the proportion of private versus public schools in Quebec did not reflect the proportion in their population. PCAP 2016 was the first administration in which provinces were asked to verify their participation rates as soon as data were received, and prior to analyses of the data. As school governance is not a stratification variable in PCAP, this discrepancy would not have been detected until later stages in the analyses in previous PCAP administrations. Statistics Canada was contracted to perform re-weighting of the data to better reflect the population in Quebec. Detailed information of the re-weighting process is presented in APPENDIX B.

***Item recoding***

Prior to the data analyses, the PCAP 2016 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 for IRT analyses. The PCAP 2016 raw assessment data included both valid and invalid responses to the test items. For a multiple-choice (MC) item, 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. 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, 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.

### *Missing data*

As is the case in other large-scale assessments, three types of data were missing from the PCAP 2016 assessment:[12]

- missing due to item sampling (not administered);
- 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:

- 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, making it impossible for the student to answer. For example, there may have been a photocopying or printing error so that the question was not legible. The not applicable code was used in only a few cases and 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.

[12] 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.

In addition, the examinee attempted no subsequent items in the remainder of the booklet.[13] 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 2016 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 they estimate theta scores they treat these items 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 (cited in de Ayala, 2009, p. 150). He argues that with the practice of ignoring omitted items, a high 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. These types of response were coded 8.

**Item analysis**

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

***Classical theory item analysis***

The objective of the classical analysis was to produce statistics for a second review of the PCAP 2016 items. For the major domain, which was reading, the first review used the field test data. The minor domains consisted of anchor items from previous administrations. Anchor items were used for the minor domains to assess the change in these items over time (or from one cohort to another) on the basis of their estimated difficulty. The mathematics items were administered in 2013 and 2010 and science items were administered in 2013. These items were

[13] 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).

field tested when mathematics and science were major domains. The statistics were reviewed in preparation for the selection of items to be included in PCAP 2016.

The classical 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 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 interscorer 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. Thus there were eight tests for the reading domain. There were also eight tests, albeit smaller, for each minor domain. For these minor domains, the placement of the items in PCAP 2016 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. 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 2016 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[14]
- 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 per cent of students selecting one of the MC detractors
- less than 10 per cent of students being awarded the score value for a CR item
- interscorer agreement of less than 70 per cent on the score value of a CR item

[14] 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.

In PCAP 2016, the classification of a new item was identified as both reading and science by content experts. It was determined that this item belongs to the reading domain through factor analysis. Correlations also show that the new item is more closely related to the reading score.

### *IRT analysis*

The IRT analysis process involved: (1) assessing the dimensionality of PCAP 2016, (2) estimating items' parameters, and (3) assessing the IRT model fits. The IRT model fits included the local item dependence (LID), the agreement between the model's mathematical function and the data, and the PCAP 2016 invariance. The process ended with assessing the differential item functioning (DIF) of the PCAP 2016 items as part of the validity evidence.

### *Assessing the dimensionality of PCAP 2016*

The PCAP 2016 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 took into account 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.

### *Item calibrations and assessing the IRT models' fit*

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 model (2PLM) to the data. It was then compared to the three-parameter 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.

### *Differential item functioning*

The differential item functioning (DIF) involved assessing the extent to which some of the PCAP 2016 items displayed different statistical proprieties (e.g., level of difficulty) for gender and language. In other words, the purpose of DIF is to determine if any item display bias towards a gender or language group. This was done through the Mantel-Haenszel (M-H) method and the Wald test implemented in IRTPRO. This test is performed in IRTPRO "with accurate item parameter error variance-covariance matrices computed using a supplemented expected maximum algorithm" (see IRTPRO technical documentation). While some of the items exhibited a DIF, this was balanced between the groups compared as an almost perfect overlap of differential test functioning made evident.

### 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 2016 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 2016. Therefore all mathematics 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.[15] The linking of PCAP 2016 was done using concurrent calibration with PCAP 2013. 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

[15] 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.

advantage of making maximum use of all the available data in estimating item parameters (Martin, Mullis, Foy, Brossman, & Stanco, 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 2016 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 at stage II to the current assessment achievement distributions.

The mathematics and the science achievement score generation for PCAP 2016 (the theta scores and the scale scores) used the item parameters estimated at this stage.

**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., reading, mathematics, 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.
- For reading subdomains, all the weighted scores were reset to a Canadian mean of 500 and standard deviation of 100. This was due to the disconnecting of links between 2007 (when reading was last the major domain) and 2010 (when the target sample changed)—the reading subdomain scores were not rescaled in 2010. The reading subdomain scores for PCAP 2016 became the new baseline for future student achievement mean comparisons.

**Standard error estimates**

The PCAP 2016 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. While the bootstrapping approach is more and more widely used, especially in research fields, it can suffer from not yielding consistent estimates if the seed and sorting of

variables change at each run. This is because there are many random samples that can be drawn from the initial sample. As a result, if someone wants to replicate the standard error for the PCAP results there is a likelihood of different results. In order to provide the reader a closer estimate of the standard errors reported in the PCAP 2016 reports, the bootstrapping seeds and the sorting variables are provided in APPENDIX C. During the bootstrap estimations, the student weights were used for analysis of student achievement.

**Presentation of the PCAP 2016 achievement results**

Summary score reports were developed at the Canadian, provincial, language, and gender levels for each of the three achievement domains. Results were provided in tabular and graphic formats and followed the pattern set out in the PCAP 2007, 2010, and 2013 public reports. Ninety-five per cent confidence intervals were calculated using the bootstrapped standard errors. T-tests were conducted for all comparisons made with the use of Bonferroni adjustments based on the number of comparisons.

# Chapter 10. Analysis of Questionnaire Data

As in previous assessments, PCAP 2016 collected background data of students, teachers, and schools. The school questionnaire was filled out by the school principles. Student questionnaire were completed on paper along with the assessment booklets, for teachers and school questionnaire, teachers and school principals had the options to complete their questionnaire on-line or on paper. The analysis of the questionnaire data included:

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

These statistical analyses were conducted for each of the three questionnaires and were reported by language. PCAP 2016 is the first 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 on-line 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. In one province, data for teacher and school questionnaires were lower than expected. Investigation revealed that the data from the paper-based questionnaires had not been collected, and so re-weighting of the teacher and school weights were necessary for that province.

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 2016 included valid and invalid responses. A response to a question was valid if only one response option was chosen but the response was considered invalid if more than one option was chosen. 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 had written responses, which required coding. Numeric variables were required for quantitative analyses of these items. These recoded items included the countries in which students and their parents were born, and the types of materials students like to read.

***Missing data***

Three types of missing data occurred in the PCAP 2016 questionnaires:

- missing responses because a student ran out of time to complete the questionnaire (not-reached); [16]
- omitted items, that is, items skipped by a student intentionally or unintentionally throughout the instrument.
- 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. They included frequency and percentage distributions for all items on categorical and Likert-type scales. The descriptive statistics also included the mean, the standard deviation, and the shape statistics (skewness).

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

**Factor analysis**

In the questionnaires, a set or block of times was used to explore specific characteristics or attitudes. Factor analysis involved performing exploratory factorial analysis (EFA) of PCAP 2016 questionnaire items by grouping the blocks of items into *factors*. Some items were recoded to match similar items for factor analyses. The resulting factor scores were scaled to the mean of 50 and standard deviation of 10 to become *index scores*. Index scores were then correlated with reading achievement, and due to the large sample size, all correlations were found to be significant, although in some cases the correlation was low. As was the case in previous PCAP assessments, only factors with correlations of .2 or above were reported in the PCAP 2016 Contextual Report. Index scores were also divided into four approximate equal quarters as follows:

- bottom quarter: below 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 PCAP 2016 questionnaires. An analysis was conducted for each questionnaire and was reported by province and by language. It primarily focused on the following items:

- mean and standard deviation
- correlations between items
- percentage of respondents with missing responses to the items
- correlations and regressions between index scores and student achievements
- regression analysis, and
- Cronbach's alpha for each index

**Group comparison analysis**

The group comparison analysis involved:

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

**Correlational analysis**

The correlational analysis included:

- computing simple correlation coefficients, also named the bivariate or the zero-order correlation, between student achievement and a background variable or an index variable.
- performing linear multiple regression analysis to predict achievement in reading 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 reading from a set of teacher- and school- related variables.
- performing linear multiple-regression analysis to predict achievement at the school level, that is, the schools mean achievement in reading from a set of teacher- and 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.

# Chapter 11. PCAP Data Sets

**Description of the data sets**

All the PCAP 2016 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. This data set can establish relationships between student performance and the contextual data. The data sets come in SPSS and Excel formats. Variables labels were set up beforehand on SPSS, so no codebook is provided for SPSS; however, a codebook is provided for the data set in Excel format.

***Student data set***

This data set includes primarily the following data:

- general information about students (student, school, and teacher identification numbers; student participation code; student use of accommodations; booklet number; each student's province and language);
- student statistical weights;
- responses to the Student Questionnaire items;
- student index scores; and
- achievement scores and performance levels for all domains and for reading subdomains.

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.[17]

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

[17] Teacher index scores were not included in the contextual report.

### *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.[18]

### *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. The number of cases in this data set will be equivalent to the number of cases in the student data set. It should be noted that depending on the type of analyses, the merged data set, which is based at the student level, may not be appropriate for all analyses. The correct weights should also be used depending on whether the level of variables of interest is based at the students, teacher, or school levels.

### Accessing the data set for research

PCAP, a pan-Canadian assessment with well-structured contextual questionnaires, afforded 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 provides 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/territorial[19] 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 here.

[18] School index scores were not included in the contextual report.
[19] No territories participated in PCAP 2016.

**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 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(ies) 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 are 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 Web site 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. Retrieved from http://www.powershow.com/view/3c4039-MmRjY/An_Introduction_to_Multidimensional_IRT_Derek_Briggs_April_powerpoint_ppt_presentation

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.

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

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

Council of Ministers of Education, Canada. (2005). *The pan-Canadian assessment program: Literature review of science assessment and test design*. Toronto: Author (unpublished report).

Council of Ministers of Education, Canada. (2011). *PCAP-2010: Pan-Canadian assessment program*. Toronto: Author.

Council of Ministers of Education, Canada (CMEC). (2016). *PCAP 2016 Assessment Framework*. Toronto: Author.

de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: 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.

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].* Welland, ON: Niagara College. Retrieved from http://noharm.software.informer.com/

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*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP11_Scaling_Achievement.pdf

Hidi, S., & Berndorff, D. (1998). Situational interest and learning. In L. Hoffmann, A. Krapp, K.A. Renniger, & J. Baumert (Eds.), *Interest and Learning*. Kiel, Germany: 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.

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 a la langue française. *Cahiers Études de Radio-Télévision*, *19,* 253–274.

Kim, S., & Kolen, M.J. (2006). Robutness 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: Zakaluk, B.L., Samuels, S.J., (eds.), *Readability, its past, present and future*. Newark, Delaware: 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.). Bruxelles: De Boeck.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264.

Lord, F.M. (1983). Maximium 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.

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

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. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_05_Chapter_2.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,* Reston, VA: Author.

O'Grady, K., Fung, K., Servage, L., & Khan, G. (2018). *PCAP 2016: Report on the pan-Canadian assessment of reading, mathematics, and science*. Toronto: Council of Ministers of Education, Canada.

Organisation for Economic Cooperation and Development (OECD) (2012). *PISA 2009 Technical Report.* Paris: PISA, OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264167872-en

Organisation for Economic Cooperation and Development (OECD) (2006). *PISA 2006: Science Competencies for Tomorrow's world.* Paris: PISA, OECD Publishing.

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.

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**

# Pan-Canadian Assessment Program

# Sample Design and Weighting

Developed for CMEC by
Statistical Consultation Group
International Cooperation and Corporate Statistical Methods Division
Statistics Canada
November 2015 (Sample Design)
November 2016 (Weighting)

**Table of Contents**

1. Introduction ..... 3
2. Sample design ..... 3
2.1 Target and survey populations ..... 3
2.2 Survey frame ..... 4
2.2.1 First stage survey frame (list of in-scope schools) ..... 5
2.2.2 Second stage survey frame (list of in-scope classes) ..... 7
2.3 Stratification ..... 8
2.4 Sample size allocation ..... 9
2.5 Sample selection ..... 11
3. Sampling weights ..... 12
4. Sampling weights non-response adjustments ..... 13
5. Bootstrap weights ..... 16
Table A – Recoded participation status ..... 19
Table B – Summary of manual school ID adjustments ..... 20
References ..... 23

## 1. Introduction

The Council of Ministers of Education Canada (CMEC) is responsible for administering the fourth iteration of the Pan-Canadian Assessment Program (PCAP). This iteration of PCAP took place in the spring of 2016 and was designed to evaluate the competencies of grade eight students in the fields of science, mathematics, and reading.

CMEC has contracted Statistics Canada to design and implement the probability sampling strategy for PCAP 2016. The purpose of this document is to detail all aspects of this sample design. More specifically, this document 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.

## 2. Sample design

The *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 subsections 2.1 to 2.5.
As mentioned above, the objective of PCAP is to assess certain competencies of grade eight students from Canada's ten provinces. Since a single list of all grade eight students in Canada does not exist, the PCAP sample was selected using a stratified, two-stage probability sampling design. This means that the sample of grade eight students was selected in two successive stages.

In the first stage, a random sample of schools (primary sampling units or PSU) was selected from a list of in-scope schools provided by CMEC. All in-scope grade eight classes within the selected schools were then enumerated. The second stage of sample selection consisted of randomly selecting one grade eight class per selected school. The grade eight classes were, therefore, the secondary sampling units (SSU). All in-scope students within the selected grade eight classes received the PCAP evaluation.

### 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.

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 is all grade eight students in the ten Canadian provinces. Schools that are not funded by the provinces are excluded from the target population. The following types of students are also excluded from the target population:

- students with functional or intellectual disabilities
- students from federal or international schools
- students who 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 excludes geographically isolated schools and schools with fewer than five grade eight students. Our analysis suggests that these exclusions account for no more than 0.7% and 2.1% of the target population at the national and provincial levels respectively. While selected schools that previously participated in other evaluations were not explicitly excluded from the coverage of this study, PCAP administrators may have chosen to replace them with selected replacement schools.

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 eight students. For this study, the responding units are the grade eight students from 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 took place, i.e., spring of 2016.

### 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 are two survey frames—one for each stage of sample selection. Subsections 2.2.1 and 2.2.2 give further detail on the survey frames for the first and second stages 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 eight students within the school, the number of grade eight 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 2.1).

**Table 2.1: Important variables for the first stage of sample selection**

| Variable | Description |
|---|---|
| JURIS_SCH_ID | School identification number |
| PROVINCE | Province |
| SCHOOLBOARD | Language of school board (English or French) |
| SCHOOL_GR8 | Number of grade eight students |
| SCHOOL_CLASSES | Number of grade eight classes |
| EXEMPTION | School exclusion flag |

Section 2.1 delineated the few types of schools that are not covered by this study. To remove these schools from the survey frame before sampling, the exemption flag was used. The following table (Table 2.2) provides documentation for the types of schools excluded from this study.

**Table 2.2: Exemption flag documentation**

| Exemption | Description | Comment |
|---|---|---|
| 1 | Special school | Excluded |
| 2 | School within other province | Excluded |
| 3, 9 | Geographically isolated school | Excluded |
| 4 | Already writing another assessment | Not excluded. May be replaced with replacement school. |
| 5, 7, 8 | Federal/international school | Excluded |
| 6 | Schools smaller than 9 students | Not excluded. SCHOOL_GR8<5 used instead. |
| 10 | Schools that are not funded | Excluded |
| 11 | Schools now closed | Excluded |

To verify the coverage of the PCAP 2015 frame, we compared the number of grade eight students reported on the frame to census population projections of 13-year-olds. We also compared these counts to those observed on the PCAP 2013 frame (Table 2.3). Considering these results, we see that the number of grade eight students on the 2016 frame aligns well with the counts from the 2013 frame, as well as with the census projections. In 2015, at the national level, we see that the PCAP frame offers 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 89.4% in Manitoba. Given these comparisons, we have confidence in the quality of the PCAP frame and conclude that it is of good quality.

**Table 2.3: Coverage of the PCAP frame**

| | Population Projections* (13-year-olds) | | Frame | | Proportion of Population | |
|---|---|---|---|---|---|---|
| **PROVINCE** | **2013** | **2015** | **2013** | **2015** | **2013** | **2015** |
| British Columbia | 47,129 | 45,848 | 46,895 | 45,576 | 99.5% | 99.4% |
| Alberta | 45,386 | 45,354 | 40,094 | 46,578 | 88.3% | 102.7% |
| Saskatchewan | 13,677 | 13,371 | 12,598 | 12,335 | 92.1% | 92.3% |
| Manitoba | 16,289 | 15,852 | 14,451 | 14,174 | 88.7% | 89.4% |
| Ontario | 153,091 | 147,322 | 145,756 | 146,765 | 95.2% | 99.6% |
| Quebec | 79,921 | 78,354 | 85,435 | 81,563 | 106.9% | 104.1% |
| New Brunswick | 8,010 | 7,543 | 7,976 | 7,327 | 99.6% | 97.1% |
| Nova Scotia | 9,785 | 9,146 | 9,792 | 8,674 | 100.1% | 94.8% |
| Prince Edward Island | 1,657 | 1,527 | 1,487 | 1,433 | 89.7% | 93.8% |
| Newfoundland and Labrador | 5,436 | 5,221 | 5,441 | 5,193 | 100.1% | 99.5% |
| **CANADA** | **380,381** | **369,538** | **369,925** | **369,618** | **97.3%** | **100.0%** |

*Source: Statistics Canada. Table 051-0001 – With data derived from estimates of population, by age, group, and sex for July 1, Canada, provinces and territories, annual (persons unless otherwise noted)

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

Once a sample of schools was selected, CMEC worked with the provinces to enumerate all grade eight classes in the selected schools. As eighth grade students may move between classes throughout the day, it was recommended that the enumeration be of grade eight homerooms or some class that every eighth grade student has to take (e.g. math or mother tongue). Table 2.4 documents the important variables required for the second stage of sample selection.

**Table 2.4: Important variables for the second stage of sample selection**

| Variable | Description |
|---|---|
| SAMPLE ID | School identification number from school sample file |
| CLASS ID | Within each school, unique class identification number |
| TEACHER NAME | Name of teacher, redundant if CLASSID is unique |
| GRADE 8 STUDENTS | Number of grade eight students within each class |

| | |
|---|---|
| EXEMPTION | Class exclusion flag |

It may be that some schools do not have well-defined classes. In this case, we recommended that all grade eight students be enumerated. It may be that some schools do not have well-defined grades. In this case, we recommended that all students with dates of birth between January 1, 2002, and December 31, 2002, be enumerated.

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 we exempt students; 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 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.

All exemptions at the class level must be approved by CMEC.

### 2.3 Stratification

*Stratification* is a means of organizing the sampling frame so that a 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.

It is important for CMEC to publish reliable statistics at the national level, at the provincial level, as well as on the language of the school systems within each province. To ensure a large enough sample within these domains, the PCAP strata are defined as the cross-classification of province by language of school board.

## 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 PCAP 2016, we started the sample size allocation process by considering the allocation strategy that was used in 2013. In examining this allocation strategy, we determined that it will perform admirably in achieving the required precision requirements. This allocation had the further advantage of being pre-approved by the provinces. For these reasons, the 2016 sample size allocation is essentially unchanged from the previous survey iteration. The table below (Table 2.5) specifies the 2016 sample size allocation. Considering these results, we see that a census of schools will be taken in 13 of 20 strata.

**Table 2.5: PCAP 2016 sample size allocation**

| | In-Scope Population | | Sample | | | |
|---|---|---|---|---|---|---|
| **Stratum** | **Schools** | **Students** | **Schools** | **Students *** | **Take all Schools** | **Take all Students** |
| **Alberta** – Anglophone | 725 | 45,622 | 150 | 2,990 | | |
| **Alberta** – Francophone | 20 | 426 | 20 | 299 | Yes | |
| **British Columbia** – Anglophone | 462 | 44,887 | 150 | 3,011 | | |
| **British Columbia** – Francophone | 12 | 262 | 12 | 171 | Yes | |
| **Manitoba** – Anglophone | 321 | 13,539 | 150 | 2,846 | | |
| **Manitoba** – Francophone | 19 | 369 | 19 | 267 | Yes | |
| **New Brunswick** – Anglophone | 83 | 5,122 | 83 | 1,516 | Yes | |
| **New Brunswick** – Francophone | 61 | 2,195 | 61 | 1,164 | Yes | |
| **Newfoundland** – Anglophone | 114 | 5,039 | 114 | 1,865 | Yes | |
| **Newfoundland** – Francophone | 3 | 23 | 3 | 23 | Yes | Yes |
| **Nova Scotia** – Anglophone | 116 | 8,333 | 116 | 2,480 | Yes | |
| **Nova Scotia** – Francophone | 11 | 329 | 11 | 184 | Yes | |
| **Ontario** – Anglophone | 2,930 | 139,230 | 150 | 2,713 | | |
| **Ontario** – Francophone | 151 | 6,367 | 125 | 2,094 | | |
| **Prince Edward Island** – Anglophone | 22 | 1,369 | 22 | 438 | Yes | |
| **Prince Edward Island** – Francophone | 3 | 58 | 3 | 58 | Yes | Yes |
| **Quebec** – Anglophone | 89 | 7,425 | 89 | 1,801 | Yes | |
| **Quebec** – Francophone | 480 | 71,909 | 150 | 3,361 | | |
| **Saskatchewan** – Anglophone | 440 | 11,992 | 186 | 3,143 | | |
| **Saskatchewan** – Francophone | 5 | 79 | 5 | 79 | Yes | Yes |
| **CANADA** | **6,067** | **364,575** | **1,619** | **30,503** | **13/20** | **3/20** |

*Expected sample size

## 2.5 Sample selection

*Sample selection* refers to the process used to obtain the survey sample from the survey population. The purpose of this subsection 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 13 of 20 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 eight 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.

The CMEC must make every effort to confirm the participation of as many sampled schools as possible. This is important in order to minimize the potential for non-response biases. After all contacts with sampled schools have been made, the CMEC will need to contact replacement schools for those sampled schools that will not participate. Each sampled school that can/will not participate should be replaced if possible by the first replacement school. Second replacement schools should only be used if both the corresponding sampled school and first replacement school do not participate. ***If the original sampled school is ineligible, or is a type of school belonging to defined school-level exclusions or has been closed, then no replacement schools should be used***. Additionally, a school with a small but sufficient number of grade eight students should not be replaced simply because the number of students might be 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, we 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. 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), three selection methods were used. For schools with more than twenty students and who were able to enumerate all grade eight classes, one grade eight class per school was selected via a Simple Random Sample (SRS). All in-scope students within the selected classes were evaluated. For schools with more than twenty students, but are unable to enumerate grade eight classes, an enumeration of all grade eight students was completed and a SRS of twenty students was selected. For schools with twenty or fewer grade eight students, a census of students within these schools was taken.

**3. Sampling weights**

Sub-section 2.5 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.

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 there are $H = 20$. Let $n_h$ be the sample size allocated to stratum $h$. Let $X_i$ be the total number of grade eight students in school $i$. Then the design weight ($d_i^{school}$) for the selected school $i$ is given as:

$$d_i^{school} = \begin{cases} \frac{\sum_{i \epsilon h} X_i}{n_h X_i}, & if\ n_h < N_h \\ 1, & if\ n_h = N_h \end{cases}.$$

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:

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

where $m_i$ is the total number of grade eight classes within school $i$. The student design weight ($d_i^{student}$) is defined as:

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

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 is necessary for the CMEC to provide Statistics Canada with the total number of in-scope students per selected class, or at the school if classes cannot be enumerated.

The CMEC, in collaboration with the schools and the provinces, conducts the PCAP evaluations. Once again, all selected students or all in-scope students within selected classes are to be evaluated. As there are four distinct test booklets, we recommend that the distribution of the booklets have a single random starting point within each province. As well, we recommend that the distribution of booklets be coordinated across schools. That is, school one starts distributing booklets at the random start and school two starts distributing booklets where school one has left off and so on. This method ensures equitable distribution of each booklet.

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

Recall the initial school weight:

$$(d_i^{school}, \{\textbf{\textit{School\_design\_weight}}\}^1):$$
$$d_i^{school} = \begin{cases} \frac{\sum_{i \epsilon h} X_i}{n_h X_i}, & if\ n_h < N_h \\ 1, & if\ n_h = N_h \end{cases}.$$

Then there is the initial school non-response adjustment factor ($nr\_initial_i^{school}$ {school_nr_adjust}):

$$nr\_initial_i^{school} = \frac{\sum_{i\ participating\ schools} d_i^{school} + \sum_{i\ non-respondent\ schools} d_i^{school}}{\sum_{i\ participating\ schools} d_i^{school}}.$$

The interim school weight ($w_i^{school}$ {***school_weight***}) is given by:

$$w_i^{school} = d_i^{school} * nr\_initial_i^{school}.$$

[1] The corresponding variable names on the files are given in { }.

A subsequent non-response adjustment was created for the school non-response that was found, often due to work-to-rule or other unplanned disruptions, with this factor ($nr\_sub_i^{school}$ {school_nr_adjust}):

$$nr\_sub_i^{school} = \frac{\sum_{i\ participating\ schools} w_i^{school} + \sum_{i\ non-respondent\ schools} w_i^{school}}{\sum_{i\ participating\ schools} w_i^{school}}.$$

With the final school weight ($w_{final}{}_i^{school}$ {***school_weight_final*** }) given by:

$$w_{final}{}_i^{school} = nr_{sub}{}_i^{school} * w_i^{school} = nr_{sub}{}_i^{school} * nr_{initial}{}_i^{school} * d_i^{school}.$$

Note that the exempted classes are not part of the adjustment. The exempted classes are given a weight that is kept to represent all the other exempted classes not found, due to the out-of-date frame. These are kept so that exemption rates can be calculated; exemption rates and participation rates are the most commonly used quality indicators for education assessments.

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 due to the answer sheet/booklet not being returned are all considered as non-respondents.[2] 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 ($nr_i^{student}$ { student_nr_adj}, which is performed within each class is expressed as:[3]

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

Using a concrete example, if there were 20 students in the class, with 10 who participated in the assessment, two who were absent, and eight who 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 ( $w_i^{*\,student}$ {***student_weight_final*** } ) is given by:

$$w_i^{*\,student} = d_i^{school} * nr_{initial}{}_i^{school} * nr_{sub}{}_i^{school} * adj_i^{class} * nr_i^{student},$$

or using the variable names:

***student_weight_final***
***= School_design_weight * school_nr_adjust * school_nr_adjust_again * class_adj * student_nr_adj.***

[2] See Table A – Recoded participation status, p. 19, for a list of participation status and associated response codes
[3] *Class* is defined here as either an actual class or grouping of (small) classes to make a *de facto* class.

Keeping with the standards of other international educational surveys, a dichotomous variable (IN_PCAP_2016) was created to indicate whether a student participated in the survey, with:

$$in\_PCAP\_2016_j = \begin{cases} 1, & if\ student\ j\ participated\ in\ the\ assessment \\ 0, & otherwise \end{cases}.$$

Therefore, a final weight variable was created for the participating students (***student_weight_final_part***) with:

***student_weight_final_part = student_weight_final * in_PCAP_2016***.

As described above, weights are derived at school and student levels. These weights are related to the student component of the survey.

There are two other components in this survey: school and teacher questionnaires. Given that not all school and/or teacher data were obtained for schools with responding students, new weights have to be derived in order to produce estimates at school and teacher levels for their respective data collected.

**School weight (for data collected at the school level)**

At student level, school weight is obtained with the following formula (as described previously):

$$w_{final}{}_i^{school} = nr_{sub}{}_i^{school} * nr_{initial}{}_i^{school} * d_i^{school}.$$

From the post-stratification applied to Quebec, the following formula was applied (see APPENDIX B):

$$w_{final}{}_i^{school\_pt} = \begin{Bmatrix} w_{final}{}_i^{school}\ for\ all\ provinces\ except\ Quebec \\ w_{final}{}_i^{school} * PST_{adj}\ for\ Quebec \end{Bmatrix}.$$

From the file provided by CMEC, all schools for which no data were obtained can be identified. The following non-response adjustment is then created:

$$nr\_sdata_i^{school} = \frac{\sum_{i\ participating\ schools} w_{final}{}_i^{school\_pt} + \sum_{i\ non-respondent\ schools} w_{final}{}_i^{school\_pt}}{\sum_{i\ participating\ schools} w_{final}{}_i^{school\_pt}}.$$

The final school weight is given by:

$$w_{final}{}_i^{school\_data} = nr\_sdata_i^{school} * w_{final}{}_i^{school\_pt}.$$

**Teacher weight (for data collected at the teacher level)**

In each school selected, the teacher (all of them if more than one) is invited to complete a questionnaire. So, the school weight and the number of classes are considered to derive the teacher weight.

As mentioned in the previous section, school weight is given by:

$$w_{final_i}^{school\_pt} = \begin{Bmatrix} w_{final_i}^{school} \; for \; all \; provinces \; except \; Quebec \\ w_{final_i}^{school} * PST_{adj} \; for \; Quebec \end{Bmatrix}.$$

So, the teacher weight is obtained by multiplying the school weight and the number of grade eight classes:

$$w_{final_i}^{teacher\_pt} = w_{final_i}^{school\_pt} * \; adj_i^{class}.$$

From the file "PCAP 2016 List of schools and teachers.xlxs" provided by CMEC, all teachers for which no data were obtained can be identified. The following non-response adjustment is then created:

$$nr\_tdata_i^{teacher} = \frac{\sum_{i \; participating \; teachers} w_{final_i}^{teacher} + \sum_{i \; non-respondent \; teachers} w_{final_i}^{teacher}}{\sum_{i \; participating \; teachers} w_{final_i}^{teacher}}.$$

The final teacher weight is given by:

$$w_{final_i}^{teacher\_data} = nr\_tdata_i^{teacher} * \; w_{final_i}^{teacher\_pt}.$$

**5. 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 (2010). 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 & Wu (1988) 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 (2012) 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*; and
- $m_k^*$ is the number of times school *k* was chosen out of the *n-1* trials.

In total, 1,000 bootstrap weights were created (bsw1,…bsw1000), with the zeroth bootstrap weight (bsw0) 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 census of classes $a_{h,k}$ = 1. See Table 5.1 for the allocations.

**Table 5.1 School allocations by census strata**

| | Schools | | |
|---|---|---|---|
| **Stratum** | **Total Number** | **Number Sampled** | **Number Participating** |
| **British Columbia** – Francophone | 12 | 12 | 12 |
| **New Brunswick** – Anglophone | 83 | 83 | 83 |
| **New Brunswick** – Francophone | 61 | 61 | 61 |
| **Nova Scotia** – Anglophone | 116 | 116 | 116 |
| **Nova Scotia** – Francophone | 11 | 11 | 11 |
| **Prince Edward Island** – Anglophone | 22 | 22 | 23 |
| **Prince Edward Island** – Francophone | 3 | 3 | 3 |
| **Saskatchewan** – Francophone | 5 | 5 | 5 |
| **"Volunteer" schools** | 3 | 3 | 3 |

The “volunteer” schools, i.e., substitution of the originally selected school with something other than the pre-selected replacement school,[4] form their own stratum, as per Kish’s (1963) recommendation for “surprises.” These students in these schools have a weight of one; they represent only themselves and no other students.

Note that given that the number of students in the class (and school) at the time of collection is likely different from the counts at the time of sample selection, the sum of the bootstrap weights (bsw1…bsw1000) for each replicate will vary from the sum of the final student weights (bsw0 = student_weight_final_part).

[4] See Table B – Summary of manual school ID adjustments, p. 20, for additional information on required adjustments.

**Table A – Recoded participation status**

| Participation Status | | | |
|---|---|---|---|
| **Participation_Code** | **Participation_Code Description** | **Grouped (Stud_Resp)** | **Stud_Resp Description** |
| 1 | Absent | NR | Non-Respondent |
| 2 | Participated during scheduled session | Part | Participant |
| 2A | Participated during scheduled session with an accommodation | Part | Participant |
| 3 | Participated during makeup session | Part | Participant |
| 4 | Exempted by the school | Excl | Excluded |
| 5 | Exempted because appropriate modifications could not be made | Excl | Excluded |
| 6 | No longer enrolled in this school/class. | Left | Left Permanently |
| 7 | Parents and/or students who do not wish to write | NR | Non-Respondent |
| 8 | Not a Grade 8 student | Excl | Excluded |
| 9 | Homeschooled students | Excl | Excluded |
| 10 | AnswerSheet and booklet were not returned. Only Questionnaire data | NR | Non-Respondent |

**Table B – Summary of manual school ID adjustments**

*Scenario 1: Renumbering of ClassIDs in Quebec*

After considerable investigative work, it was discovered that classes from Quebec had been submitted to Statistics Canada with class number of 00, but later in the process had been subject to conversion disallowing the use of 00 as a class number. Fortunately, the process to renumber the class numbers was discovered, and was found to be consistent. Essentially, class number 00 became the next available class number within the school. For example, if the school had only one class with a class number of 00, then the class number became 01. If the school had three classes, with class numbers of 00, 01, and 02, then the classes with class number 01 and 02 would remain the same, and the class with class number 00 would become class number 03.

The table below summarized the selected classes in Quebec in which the numbering of the class number was altered from 00. Note that *ClassID* is the concatenation of the four-digit school id with the two-digit class number.

| ClassID | |
|---|---|
| **Originally Submitted** | **Post-Collection** |
| 432500 | 432502 |
| 232800 | 232801 |
| 633100 | 633102 |
| 633700 | 633706 |
| 233900 | 233901 |
| 234100 | 234101 |
| 234200 | 234201 |
| 634300 | 634305 |
| 234700 | 234701 |
| 635600 | 635601 |
| 435700 | 435701 |
| 635800 | 635801 |
| 235900 | 235901 |
| 236000 | 236001 |
| 436100 | 436101 |
| 436200 | 436201 |
| 436500 | 436504 |
| 437100 | 437101 |
| 437200 | 437201 |
| 637400 | 637401 |
| 637700 | 637701 |
| 437800 | 437801 |
| 437900 | 437901 |

| ClassID | |
|---|---|
| **Originally Submitted** | **Post-Collection** |
| 438000 | 438001 |
| 239100 | 239101 |
| 239300 | 239301 |
| 239500 | 239501 |
| 439600 | 439601 |
| 439700 | 439701 |
| 440000 | 440001 |
| 440300 | 440301 |
| 640400 | 640401 |
| 440500 | 440501 |
| 440700 | 440701 |
| 241006 | 241806 |
| 241900 | 241901 |
| 642200 | 642201 |
| 442300 | 442301 |
| 642500 | 642501 |

*Scenario 2: Originally misclassified schools*

Using the following explanation submitted by CMEC: "At the beginning, there was one school with three classes. Then we found out that the school has two distinct sections (English/French) and therefore has to be treated as two different entities. *School part 1 with two classes* and *School part 2 with 1 classes*. The sample was redrawn. The original ClassID 220202 needs to change to 438401." This change to the school is documented below:

| ClassID | |
|---|---|
| **Originally Submitted** | **Post-Collection** |
| 220202 | 438401 |

*Scenario 3: Schools originally exempted, later found to have in-scope students*

At the time of listing, some schools had no in-scope students, or selected classes with no in-scope students. Although these schools had been removed from the weighting information at Statistics Canada, their weighting information was re-instated after classes were found on the participating students' database. These schools are documented below:

| ClassID | |
|---|---|
| **Originally Submitted** | **Post-Collection** |
| 107301 | 107301 |
| 132902 | 132902 |
| 217702 | 217702 |

*Scenario 4: "Volunteer" schools –school substitution*

The "volunteer" schools, i.e., substitution of the originally selected school with something other than the pre-selected replacement schools, form their own stratum, as per Kish's (1963) recommendation for "surprises." These students in these schools have a weight of one; they represent only themselves and no other students. These schools are documented below:

| **Originally Submitted ClassID** | **Post-Collection STRATUMID** |
|---|---|
| 123202 | 9991_1 |
| 452301 | 9994_1 |
| 452301 | 9995_1 |

**References**

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

Kish, L. (1963) *Survey Sampling*, John Wiley & Sons, Inc.

Lohr, S. (2010). *Sampling: Design and Analysis: 2nd Edition,* USA: Duxbury Press.

Rao, J.N.K. & Wu, C.F.J. (1988). "Resampling inference with complex survey data", *Journal American Statistical Association*, 83, 231–241.

**APPENDIX B**

## Adjustment for Quebec Weights (Post-Stratification)
## November 2017

For all provinces, the final weights produced were derived based on the sampling design in place and a nonresponse adjustment. The stratification variables used were the province and the language. For the province of Quebec, we observed that the distribution of public and private respondent schools was quite different than what it appears on the sampling frame for francophone schools (Table 1). Unfortunately, this situation decreased the precision of the estimates produced.

**Table 1: Distribution of public and private schools: sampling frame vs sample**

| Stratum | Level | Public Schools | | Private Schools | |
|---|---|---|---|---|---|
| | | Frame | Sample | Frame | Sample |
| **Quebec** – Francophone | School | 75% | 47% | 25% | 53% |
| **Quebec** – Francophone | Student | 79% | 38% | 21% | 62% |
| **Quebec** – Anglophone | School | 75% | 77% | 25% | 23% |
| **Quebec** – Anglophone | Student | 87% | 88% | 13% | 12% |

Information about public and private schools for the survey population is available on the sampling frame (Table 2). This information can be incorporated in the weighting process and, then, improve the precision of the estimates.

Post-stratification is used to adjust the survey's weights when the auxiliary data are in the form of counts. For the province of Quebec, the counts were obtained from the sampling frame.

**Table 2: Sampling frame counts**

| Stratum | Level | Population Counts | |
|---|---|---|---|
| | | Public Schools | Private Schools |
| **Quebec** – Francophone | School | 361 | 119 |
| **Quebec** – Francophone | Student | 56,739 | 15,170 |
| **Quebec** – Anglophone | School | 67 | 22 |
| **Quebec** – Anglophone | Student | 6,457 | 968 |

Given that auxiliary data are available at both levels, post-stratification is used for both of them. For each level, the post-stratified adjustment factor is computed for each post-stratum (public and private) (Table 3). At the school level, this factor corresponds to the ratio of the number of schools in the post-stratum to the estimated number of schools in the post-stratum, which is estimated using the original school final weights. Similarly for the student level, this factor corresponds to the ratio of the number of students in the post-stratum to the estimated number of students in the post-stratum, which is estimated using the original student final weights.

The estimates of the number of public and private schools (students) obtained after applying this post-stratified adjustment factors to the correspondent original final weights are now consistent with the sampling frame counts.

**Table 3: Post-stratified adjustment factors used**

| Stratum | Level | Post-Stratified Adjustment Factors ($PST_{adj}$) | |
|---|---|---|---|
| | | **Public Schools** | **Private Schools** |
| **Quebec** – Francophone | School | 1,5673 | 0,4678 |
| **Quebec** – Francophone | Student | 2,8467 | 0,4738 |
| **Quebec** – Anglophone | School | 1,0372 | 1,1419 |
| **Quebec** – Anglophone | Student | 1,0357 | 1,134 |

## 1. School weight

At student level, school weight is obtained with the following formula (see APPENDIX A: Pan-Canadian Assessment Program – Sample Design and Weighting):

$$w_{final}{}_i^{school} = nr_{sub}{}_i^{school} * \ nr_{initial}{}_i^{school} * d_i^{school}.$$

To take into account the previous post-stratification, new final school weight is given by:

$$w_{final}{}_i^{school\_pt} = \begin{Bmatrix} w_{final}{}_i^{school} \ for\ all\ provinces\ except\ Quebec \\ w_{final}{}_i^{school} * PST_{adj}\ for\ Quebec \end{Bmatrix}.$$

## 2. Student weight

At student level, school weight is obtained with the following formula (see APPENDIX A: Pan-Canadian Assessment Program – Sample Design and Weighting):

$$w_i^{*\,student} = d_i^{school} * nr_{initial}{}_i^{school} * nr_{sub}{}_i^{school} * adj_i^{class} * nr_i^{student}.$$

To take into account the previous post-stratification, new final student weight is given by:

$$w_{final}{}_i^{student\_pt} = \begin{Bmatrix} w_i^{*\,student}\ for\ all\ provinces\ except\ Quebec \\ w_i^{*\,student} * PST_{adj}\ for\ Quebec \end{Bmatrix}.$$

## PCAP 2016: Seeds and Sorting Variables for Bootstrapping

The following tables provided the seeds and the sorting variables of the data prior to bootstrapping. If multiple variables are listed, the data were sorted in the order of variables listed. The SAS procedure *proc surveyselect* was used. The method of sampling was *unrestricted random sampling* (method=urs), with 200 replications for all estimations.

**For the PCAP 2016 Public Report**

| Estimation | | Seed | Sorting Variable |
|---|---|---|---|
| Mean of all domain scores: Reading, Mathematics, Science | Canada overall | 30459584 | STUDENTID |
| | Canada by gender | 03951107 | |
| | Canada by language of school system | 1024130 | |
| | By province | 110042 | |
| | By province by gender | 210342 | |
| | By province by language of school system | 310042 | PROVINCE, LANGUAGE_SCHOOL |
| Mean of all reading subdomain scores | Canada overall | 30459584 | STUDENTID |
| | Canada by gender | 03951107 | |
| | Canada by language of school system | 10302419 | |
| | By province | 302110 | |
| | By province by gender | 210342 | |
| | Province by language of school system | 310042 | PROVINCE, LANGUAGE_SCHOOL |
| Percentage distributions of reading proficiency levels | Canada overall | 30459584 | STUDENTID |
| | Canada by gender | 03951107 | GENDER, STUDENTID |
| | Canada by language of school system | 1024130 | LANGUAGE_SCHOOL, STUDENTID |
| | By province | 302110 | STUDENTID |
| | By province by gender | 210342 | PROVINCE, GENDER, STUDENTID |
| | By province by language | 210342 | PROVINCE, LANGUAGE_SCHOOL, STUDENTID |

**For the PCAP 2016 Contextual Report**

| Student-Level Analyses | | | |
|---|---|---|---|
| **Estimation** | | **Seed** | **Sorting Variable** |
| Mean scores | Student first language by language of school system | 310042 | LANGUAGE_SCHOOL |
| | Student second language program by language of school system | 310042 | LANGUAGE_SCHOOL |
| | Other student variables | 03951107 | STUDENTID |
| Mean of indices | Canada overall | 30459584 | STUDENTID |
| | Canada by gender or Canada by language | 1024130 | STUDENTID |
| | By province by gender | 110042 | STUDENTID |
| | By province by language | 310042 | PROVINCE, LANGUAGE_SCHOOL |
| Mean by quarters of indices | All | 03951107 | STUDENTID |

| Teacher-Level Analyses | | | |
|---|---|---|---|
| **Estimation** | | **Seed** | **Sorting Variable** |
| Student mean (mean of mean) | All teacher variables | 03951107 | TEACHERID |
| Mean of indices | Canada overall | 30459584 | TEACHERID |
| | Canada by language of school system | 1024130 | |
| | By province | 110042 | |
| | By province by language of school system | 310042 | PROVINCE, LANGUAGE_SCHOOL |
| Mean by quarters of indices | All | 03951107 | TEACHERID |