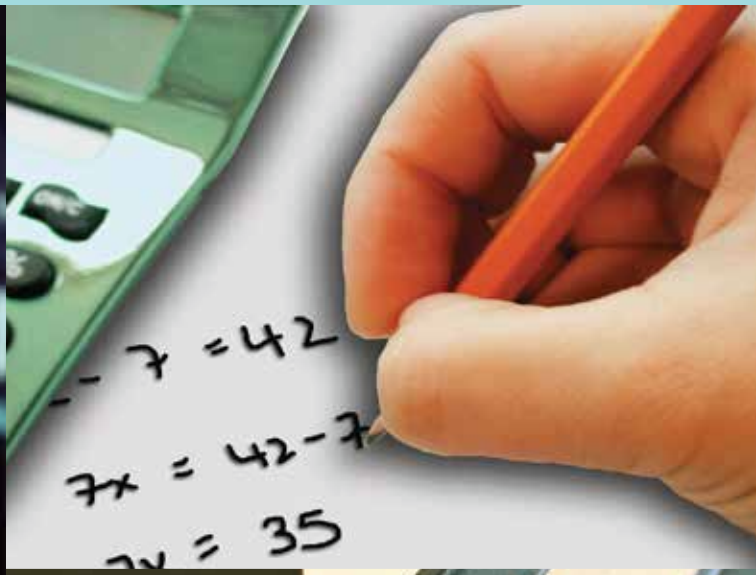
Pan-Canadian Assessment Program

PCAP 2013

Technical Report



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

Pan-Canadian Assessment Program

PCAP 2013

Technical Report

Authors

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

Koffi Houme, 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 jurisdictional 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 jurisdictions.[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,” and “jurisdictions” includes participating “provinces” and “territories.”

[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 9
Context 9
Pan-Canadian assessment 9
Participation 10
Administration time 10
PCAP in both official languages 10

Chapter 2. Design and Development of the Assessment 11
Assessment design 11
General design of the science assessment 11
Item format and item type 12
Selected-response items 12
Constructed-response items 12
Contextualized embedded attitude items 13
PCAP Science Assessment Framework 13
Describing the domain 14
Organization of the domain 14
Competencies 14
Sub-domains 16
Attitudes 18
Table of specifications 18
PCAP Reading Assessment Framework 19
Describing the domain 19
Organization of the domain 20
Linking the 2007, 2010, and 2013 reading assessments 20
PCAP Mathematics Assessment Framework 21
Describing the domain 21
Organization of the domain 22
Linking of the 2010 and 2013 mathematics assessments 23
Working groups 23
Item Development 23
Item translation and review 24
Editing and verification of the assessment items 25
Translating and comparing items in English and French 25
Readability of PCAP contexts and items 26
Determining readability 26
Readability test results 27
Editing for language and style 28
Scientific editing 28
Psychometric editing 29
Item approval by the jurisdictions 29

Chapter 3. Development of the Contextual Questionnaires 30
Initial questionnaire framework and guiding principles 30
Core questions 31
Gender differences 31
Time allocation and use 32
Science teacher efficacy and beliefs 32
Assessment 32
Accommodations (adaptations) and modifications 33
Attitudes and motivations 33
Student learning strategies 33
Teaching strategies 33
Opportunity to learn 34
Item types 34
Contextual questionnaires 34
Student Questionnaire 34
Teacher Questionnaire 35
School Questionnaire 35

Chapter 4. Sampling Procedures 37
Sampling plan 37
Sample sizes 39
Use of multiple assessment booklets 39
Existence of several sampling specifications 40
Choice of schools 42
Databases on the schools 42
Selection of schools 42
Exclusion of schools 43
Selection of students 43

Chapter 5. Field Testing 45
Item-selection working group—field study 45
Assessment booklets 45
Item-scoring session 45
Data capture 45
Data analysis 46
Item-selection working group — main study 46
Review of the assessment framework 47

Chapter 6. Main Study 48
Assessment booklets 48
Reviewing the assessment material 48
Printing the assessment booklets 48
Checking documents 48
Letter sent to parents/guardians of students 49
Administration procedures 49

Assessment site 49
Administering the assessment 49
Students with special needs 50
Questionnaires for the principal and teachers 51
Participation and exemption of students from the assessment 51
Organizing a makeup session 52
Returning assessment materials 53
Scoring session 53
Bundling booklets 53
Scoring sheets 53
Scorers' manual 54
Coding guide 54
Scorer leaders 54
Table leaders 54
Scorer training 54
Scoring reliability 55
Reliability reviews 55
Inter-rater reliability (double scoring) 57
Trend reliability 57
Multiple scoring 57
Reports and feedback 58
Jurisdictional coordinator's report 58
School Coordinator's Report 59
Scorer feedback forms 61

Chapter 7. Setting a Performance Standard 63
Standard-setting sessions 63
Selection of an expert panel 63
Preliminary performance-level descriptors 64
Security of materials 64
The Bookmark procedure 64
Standard-setting procedure 65
Performance-level descriptors 67

Chapter 8. Processing PCAP Data 68
Data gathering 68
Data capture 68
Data entry quality control 68
Data cleaning 69
General recoding 69
Review of the sampling data 69
Final review of the data and preparing the database 69

Chapter 9. Analysis of Achievement Data 71
Preliminary analysis 71
Data screening 71
Item recoding 71
Missing data 72
Not-administered items 72
Not applicable items 73
Not-reached items 73
Omitted items 73
Invalid response 73
Item analysis 74
Classical theory item analysis 74
Item difficulty 74
Item discrimination 74
Specific statistics for MC items 75
Specific statistics for CR items 75
Examining for missing data 75
Reliability of the PCAP 2013 assessment 75
Problematic items 75
IRT analysis 76
Assessing the dimensionality of PCAP 2013 76
Item calibration 76
Assessing the IRT models' fit 77
Differential item functioning 77
Linking and equating the minor domains with previous assessments 78
Test Functioning 78
Achievement score generation and scaling 79
Standard error estimates 79
Presentation of the PCAP 2013 achievement results 79

Chapter 10. Analysis of Questionnaire Data 80
Preliminary analysis 80
Data screening 80
Item recoding 80
Missing data 81
Descriptive statistics 81
Factor analysis 81
Item analysis: Classical theory item analysis 81
Group comparison analysis 82
Correlational analysis 82

Chapter 11. PCAP Databases 83
Description of the databases 83
Student database 83
Teacher database 83
School database 83
Merged database — Student/teacher/school 84
Accessing the database for research 84
Terms and conditions 84
Contact information 85

References 86

List of Tables

Chapter 2. Design and Development of the Assessment

TABLE 2.1 Number of clusters, scenarios, and items by domain and booklet 11
TABLE 2.2 Distribution of items by type and assessment domain..... 13
TABLE 2.3 Percentages allocated to competencies and sub-domains in PCAP 2013 Science 18
TABLE 2.4 Number of reading items by sub-domain and by booklet 20
TABLE 2.5 Number of mathematics items by sub-domain and by booklet 22
TABLE 2.6 Comparison of readability of PCAP science contexts and items..... 27

Chapter 4. Sampling Procedures

TABLE 4.1 Number of students in Grade 8/Secondary II, by population..... 38
TABLE 4.2 Estimating the size of a sample..... 39
Table 4.3 Structure of the assessment booklets 39
TABLE 4.4 Distribution of populations, by first-level and second-level sampling 41
TABLE 4.5 Sample size parameters 42

Chapter 6. Main Study

TABLE 6.1 Reliability review results for science scoring group 1..... 56
TABLE 6.2 Reliability review results for science scoring group 2..... 56
TABLE 6.3 Reliability review results for reading..... 56
TABLE 6.4 Reliability review results for mathematics..... 57
TABLE 6.5 Overall agreement between coders for double scoring 57

Chapter 7. Setting a Performance Standard

TABLE 7.1 Composition of the PCAP test booklets for science items..... 65
TABLE 7.2 Distribution of students by performance level in science 67

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, and 2012 Programme for International Student Assessment (PISA) (involving over 60 countries), the 2011 Progress in International Reading Literacy Study (PIRLS) (involving approximately 60 countries), and the 2013 International Computer and Information Literacy Study (ICILS) (involving approximately 20 countries). Individual jurisdictions have participated in various achievement studies, such as the Trends in International Mathematics and Science Study (TIMSS). Most jurisdictions also conduct their own evaluations of students at different stages in their schooling. To examine the teacher context, some jurisdictions 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.

School programs and curricula vary from jurisdiction to jurisdiction across the country, so comparing results from these programs is a complex task. However, young Canadians in different jurisdictions 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 jurisdictional assessments with comparative Canada-wide data on the achievement levels attained by Grade 8/Secondary II students across the country.

Information gathered by each assessment has given ministers of education a basis for examining curricula and other aspects of the school systems. 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.

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.

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

Participation

Ten provinces in Canada participated in PCAP 2013. Although Northwest Territories and Yukon have previously participated in either SAIP and/or PCAP, no Canadian territories participated in this last cycle of PCAP.

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, mathematics, and science classes. Teacher and School Questionnaires were also administered to obtain a more holistic view of the Canadian education system.

PCAP in both official languages

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 PCAP field testing and administration 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 jurisdictions 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 and Newfoundland and Labrador were not indicated in the results; however, they were included in the calculations for the overall results in those jurisdictions. Although the Saskatchewan francophone sample was also very small with only 97 students, it represented 85 per cent of the Saskatchewan Grade 8 francophone population. Reporting of data for this population was approved by the Saskatchewan Ministry of Education.

Chapter 2. Design and Development of the Assessment

This chapter describes the test design and development process for PCAP 2013. How the assessment instrument was designed is crucial to ensure that the items in the assessment booklets correctly assess the skills of Grade 8/Secondary II students across Canada. The following sections present the various stages of the assessment booklets' development for the three subjects assessed by PCAP 2013: science, reading, and mathematics. They therefore provide more detailed information on the test design and format, assessment frameworks, working groups, and how items were drafted and edited.

Assessment design

The PCAP assessment, a paper-and-pencil test, covers three assessment domains: reading, mathematics, and science. Science was the major domain, while reading and mathematics were minor domains in the 2013 study. Just as with the Programme for International Student Assessment (PISA), the focus changes with each assessment, so science will become a minor domain and reading the major domain in the next PCAP study in 2016.

For the PCAP assessment, eight clusters of science assessment units were distributed within four booklets so that each booklet contained two clusters of science items, one reading cluster, and one mathematics cluster. The four booklets were randomly and equally distributed to students within a single class. Thus, every student completed two of the eight clusters of science 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 2.1 shows the distribution of the clusters, contexts (or scenarios), and items for science, reading, and mathematics across the four booklets.

TABLE 2.1 Number of clusters, scenarios, and items by domain and booklet

	Science			Reading			Mathematics		
	Clusters	Scenarios	Items	Clusters	Scenarios	Items	Clusters	Scenarios	Items
Booklet 1	2	10	25	1	2	8	1	2	7
Booklet 2	2	10	24	1	3	8	1	3	8
Booklet 3	2	8	24	1	3	8	1	3	9
Booklet 4	2	9	24	1	3	8	1	2	9

General design of the science assessment

For PCAP assessment purposes, the domain of science is divided into three competencies (science inquiry, problem solving, and scientific reasoning); four sub-domains (nature of science, life science, physical science, and Earth science); and attitudes, within a given context. Since PCAP Science assesses scientific literacy, each assessment item is coded to both a competency and a sub-domain. Attitude items are embedded within contexts.

The competencies are interwoven throughout the sub-domains of the science assessment because they encompass the means by which students respond to the demands of a particular challenge. They reflect the current Grade 8/Secondary II science curricula for students in Canadian jurisdictions, as well as the foundation statements in the *Common Framework of Science Learning Outcomes, K to 12: Pan-Canadian Protocol for Collaboration on School Curriculum* (CMEC, 1997).[3]

Each assessment unit presents a context followed by a series of related items. The contexts chosen for assessment units were intended to captivate the interests of Canadian Grade 8/ Secondary II students and, therefore, to increase their motivation in writing the test. Contexts were introduced with an opening situation that could be in the form of a brief narrative and could include tables, charts, graphs, or diagrams. Developers of the assessment items ensured that the contexts were developmentally appropriate and not culturally or geographically dependent.

Each booklet was composed of eight to ten assessment units that together spanned each of the competencies and sub-domains. Each unit included a scenario and between one and six items. The science units were organized into eight groups or clusters.

Item format and item type

PCAP item developers selected item types that were most appropriate to what was asked. These included selected-response and constructed-response items. The test contained approximately 75 per cent selected-response and 25 per cent constructed-response items. Embedded attitude questions made up 6 per cent of the assessment.

Selected-response items

Selected-response (SR) items presented a number of responses from which the student had to choose. They included multiple-choice, check boxes, true-or-false statements, and yes–no observations. All multiple-choice items consisted of a stem statement with four choices, one of which was the correct answer while the other three were logical distractors.

Constructed-response items

Constructed-response (CR) items required students to provide a written response that could range from short phrases or two to three sentences to several paragraphs in the case of extended constructed-response items. They could also ask the student to create tables or graphs, sketch diagrams, or design experiments. PCAP Science included constructed-response items that were open-ended and measured higher-order cognitive skills and content knowledge.

[3] For updated science curricula, please visit official jurisdictional Web sites.

The inclusion of constructed-response items reflected good assessment practice —different assessment formats were required, depending on what students were expected to demonstrate. Constructed-response items allowed for partial credit, which is an important aspect when assessing process skills or for items requiring multiple steps.

Table 2.2 shows the distribution of the items types for science, reading, and mathematics across the four booklets.

TABLE 2.2 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
Science	18	7	18	6	17	7	17	7
Reading	6	2	6	2	6	2	6	2
Mathematics	3	4	5	3	5	4	5	4

Contextualized embedded attitude items

The vast majority of Canadian jurisdictions include the development of positive attitudes as an important component of science teaching and learning. This is mirrored in PCAP Science, which gathers data about students' attitudes using both contextualized embedded attitude items and a student questionnaire. Data about students' attitudes both in context (within the test) and out of context (within the questionnaire) provide information about whether attitudes vary between these two approaches and how attitude affects achievement. Hidi and Berndoff (1998) argue that although situational interest can have an important effect on both cognitive and motivational functioning, investigations into its role remain "haphazard and scattered." By using both contextualized attitude items and a student questionnaire, PCAP Science could provide data to further this area of research.

PCAP Science contained sufficient items about attitude to prepare a reliable scale; however, responses to the attitude items were not included in the overall score of scientific literacy. Nevertheless, they might provide an important component in profiling student scientific literacy, a topic to be explored in a forthcoming issue of *Assessment Matters!* (a publication available on the CMEC Web site).

PCAP Science Assessment Framework

The PCAP Science Assessment Framework was developed by consultants, pan-Canadian assessment coordinators, educators who were experts in science, and policy-makers in all jurisdictions.[4] It was informed by the curriculum objectives, goals, and outcomes of the participating populations. As well, it reflects current research findings and best practices in science learning that align with international trends

[4] *PCAP's Science Assessment Framework* is available at http://www.cmec.ca/PCAP_Science_Assessment_Framework

Describing the domain

A literature review of Canadian Grade 8/Secondary II science curricula conducted in preparation for PCAP (CMEC, 2005) clearly identifies scientific literacy as the goal of science education in all Canadian jurisdictions. For this assessment's purpose, the *PCAP Science Assessment Framework* defines scientific literacy as a student's evolving competencies in understanding the nature of science using science-related attitudes, skills, and knowledge to conduct inquiries, to solve problems, and to reason scientifically to understand and make evidence-based decisions about science-related issues.

The scope of the assessment is limited to those concepts and skills encountered and used in the courses of study of most Grade 8/Secondary II students in Canada. Although it is based on the programs taught to Canadian Grade 8/Secondary II students, this assessment is not a comprehensive assessment of all concepts and skills that a particular system expects students at this level to master. It aims to provide the jurisdictions with data to inform educational policy. It is not designed to identify the strengths or weaknesses of individual students, schools, districts, or regions.

Organization of the domain

Competencies

An understanding of science is important for young people to be able to participate in society and to recognize that science and technology affects their lives. When students are engaged with the competencies of science inquiry, problem solving, and scientific reasoning, they develop scientific literacy. PCAP Science places a priority on being able to assess these competencies.

Science inquiry involves understanding how inquiries are conducted in science to provide evidence-based explanations of natural phenomena.

Science inquiry requires students to address or develop questions about the nature of things, involving broad explorations as well as focused investigations (CMEC, 1997). From the students' perspective it involves how they focus on the "why" and "how" of science.

The PCAP assessment of students' ability to use scientific practices provides evidence that they can:

- formulate hypotheses;
- make observations;
- design and conduct investigations;
- organize and communicate information;
- analyze and interpret data (e.g., use graphs and tables);
- apply the results of scientific investigations;
- select alternative conclusions in relation to the evidence presented;
- provide reasons for conclusions based on the evidence provided; and

- identify assumptions made in reaching the conclusion.

Problem solving is using scientific knowledge and skills to solve problems in social and environmental contexts.

Problem solving requires students to seek answers to practical problems requiring the application of their science knowledge in new ways (CMEC, 1997). Students demonstrate this competency by applying their knowledge of science, their skills, and their understanding of the nature of science to solve science-related problems. Part of the process includes problem finding and problem shaping where a problem is defined as the desire to reach a definite goal.

The PCAP assessment of students' ability to solve problems provides evidence that they can:

- define the problem;
- formulate questions;
- communicate the goals related to the problem;
- solve problems by recognizing scientific ideas;
- select appropriate solutions in relation to an identified problem;
- verify and interpret results (communicate, reflect);
- generalize solutions (recognize and apply science in contexts not typically thought of as scientific);
- provide reasons for the solution and how it meets the criteria to solve the problem;
- identify assumptions made in solving the problem; and
- show an awareness of sustainable development and stewardship when addressing problems.

Scientific reasoning involves being able to reason scientifically and make connections by applying scientific knowledge and skills to make decisions and address issues involving science, technology, society, and the environment.

Scientific reasoning involves a comparison, rationalization, or reasoning from the student in relation to an existing theory or frame of reference. Students demonstrate this competency by applying their knowledge of science, their skills, and their understanding of the nature of science to make informed, evidence-based decisions. They draw conclusions or make comparisons to an existing frame of reference or perspective. Students identify questions or issues and pursue science knowledge that will inform the question or issue.

The PCAP assessment of students' ability to reason scientifically provides evidence that they can:

- recognize patterns;
- develop plausible arguments;
- verify conclusions;
- judge the validity of arguments;
- construct valid arguments and explanations from evidence;
- connect scientific ideas to produce a coherent whole;
- use reasoning to make an informed decision for a particular issue in relation to the evidence;
- use reasoning to understand a science-related issue;
- provide reasons for the decision based on the evidence provided;
- identify assumptions and limitations of the chosen decision for that issue;
- develop and use models;
- show respect and support for evidence-based knowledge; and
- display an interest in and an awareness of science-related issues.

For each competency, students are assessed on their understanding and ability to critique the practices and processes related to these competencies.

Sub-domains

The four sub-domains targeted by PCAP Science are aligned with pan-Canadian science curricula for all participating populations and with foundation statements for scientific literacy in Canada (CMEC, 1997). The four sub-domains are nature of science, life science, physical science, and Earth science.

Nature of science

PCAP defines the nature of science as involving an understanding of the nature of scientific knowledge and the processes by which that knowledge develops. Science provides a way of thinking and learning about the biological and physical world based on observation, experimentation, and evidence. Science builds upon past discoveries. Theories and knowledge are continually tested, modified, and improved as new knowledge and theories supersede existing ones. Scientific debate on new observations and hypotheses is used to challenge, share, and evaluate data through peer interaction and dissemination of information through written publications and presentations. According to Fensham and Harlen (1999), by developing students' abilities to relate evidence to conclusions and to distinguish opinion from evidence-based statements, science education promotes a deeper public understanding of science and an appreciation of evidence-based decision making, which is an important component of scientific literacy.

The PCAP assessment of students' understanding of the nature of science provides evidence that they can:

- understand the relationship among collecting evidence, finding relationships, and proposing explanations in the development of scientific knowledge;
- distinguish between processes and terminology that are scientific and those that are not;
- describe the processes of science inquiry and problem solving in evidence-based decision making;
- distinguish between qualitative and quantitative data;
- identify characteristics of measurement (e.g., replicability, variation, accuracy/precision in equipment and procedures);
- distinguish between various types of scientific explanations (e.g., hypothesis, theory, model, law);
- give examples of scientific principles that have resulted in the development of technologies; and
- demonstrate scientific literacy with respect to nature-of-science issues.

The sub-domains of life science, physical science, and Earth science are assessed through the following descriptors. (Although these descriptors reflect the commonalities of pan-Canadian curricula, they are not intended to constitute an exhaustive list.)

Life science

- Explain and compare processes that are responsible for the maintenance of an organism's life.
- Describe the characteristics and needs of living things.
- Distinguish between cells and cell components.
- Describe the function and interdependence of systems related to inputs and outputs of energy, nutrients, and waste.
- Demonstrate scientific literacy with respect to life science issues.

Physical science

- Describe the properties and components of matter and explain interactions between those components (e.g., states of matter [i.e., solids, liquids, and gases]; properties and changes of matter; particle theory; mass and volume).
- Demonstrate scientific literacy with respect to physical science issues.

Earth science

- Explain how water is a resource for society.
- Explain patterns of change and their effects on water resources on Earth (e.g., water distribution; weather; weathering and erosion; effect of water on regional climates).
- Demonstrate scientific literacy with respect to Earth science issues.

Although understanding the interrelationships between science and technology is an important part of developing scientific literacy, PCAP Science is not designed to assess the technological literacy of students writing this assessment.

Attitudes

Attitudes toward science determine students' interest in pursuing scientific careers (Osborne, Simon, & Collins, 2003). Since new scientific knowledge is essential for economic growth, students' attitudes toward science are a subject of societal concern and debate in many countries (OECD, 2006).

To analyze students' attitudes, PCAP Science assesses:

- interest in and awareness of science-related issues;
- respect and support for evidence-based knowledge; and
- awareness of sustainable development and stewardship.

Table of specifications

A table of specifications is a guide for assessment that indicates the emphasis placed on the measurement of students' understandings within the various learning competencies and sub-domains outlined above. It also reflects the degree of curricular commonality among Canadian jurisdictions. Although a higher proportion of problem-solving items were anticipated in the assessment, many items developed for this competency were found to be too difficult for students in the field tests and so could not be moved forward to the test's main administration. Table 2.3 summarizes the percentages devoted to each competency and sub-domain in the PCAP 2013 assessment.

TABLE 2.3 Percentages allocated to competencies and sub-domains in PCAP 2013 Science

Competencies	
Science inquiry	34%
Problem solving	12%
Scientific reasoning	54%

Sub-domains	
Nature of science	34%
Life science	25%
Physical science	25%
Earth science	16%

PCAP Reading Assessment Framework

The reading framework statement for PCAP 2013 has not been altered from how reading performance was defined in the 2007 assessment, where reading was the major domain.[5] This enables comparisons over time between the three cohorts.

Describing the domain

According to curricula across Canada, reading is a dynamic, interactive process whereby the reader constructs meaning from texts. The process of reading effectively involves the interaction of reader, text, purpose, and context before, during, and after reading.

The reader

To make meaning of a text, readers have to make a connection between what is in the text and what they know or bring to it. Readers' personal experiences, real or vicarious, allow a greater or lesser access to the content and forms of what they read. Knowledge of language, facility with language strategies, and knowledge of how language works in print affect the student's construction of meaning in the text.

The text

Writers produce texts for a variety of purposes and use a variety of forms. Many traditional genres have been combined or used in novel ways. Students must read a variety of texts such as those generally considered fiction and those considered non-fiction. Within that range, texts have different degrees of complexity in structure, vocabulary, syntax, organization, ideas, rhetorical devices, and subject matter. To read these forms or types successfully, students need to recognize how these genres function in different situations.

The reader's purpose

The purpose of the reading activity affects the reader's construction of meaning. Students read texts for a variety of reasons, ranging from the pleasure they take in the text's content and style to the practical information or point of view they acquire from engaging with it. Whereas particular genres are often considered aesthetic or pragmatic in intention, the reader's purpose may differ from that intent. For example, social studies students may be required to read a novel to develop knowledge of a particular culture, era, or event.

The context

Context is important in any reading act because it affects the stance the reader takes toward the printed word. Context refers specifically to the physical, emotional, social, and institutional environment at the time of reading. Any meaning constructed by a reader is a reflection of the

[5] PCAP's Reading Assessment Framework is available at http://www.cmec.ca/PCAP_Reading_Assessment_Framework

social and cultural environment in which the reader lives and reads. Peers, family, and community values affect the stance readers take as they engage with text.

Organization of the domain

In light of the interactive process linking the reader, text, purpose, and context, this assessment of the reading domain considers the reader's engagement with the text and his or her response to it. Language arts curricula across Canada identify comprehension, interpretation, and response and reflection as major organizing aspects of reading literacy. This assessment examines three sub-domains of the integrated process of reading: comprehension, interpretation, and response to text (which includes response and reflection).

Comprehension: Students understand the explicit and implicit information provided by the text. In particular they understand the vocabulary, parts, elements, and events of the text.

Interpretation: Students make meaning by analyzing and synthesizing the parts/elements/ events to develop a broader perspective and/or meaning for the text. They may identify theme/thesis and support that with references to details, events, symbols, patterns, and/or text features.

Response to text: In responding, the readers engage with the text in many ways: by making personal connections between aspects of the text and their own real/vicarious/prior experiences, knowledge, values, and/or points of view; by responding emotionally to central ideas or aspects of the text; and/or by taking an evaluative stance about the quality or value of the text, possibly in relation to other texts and/or social or cultural factors.

Table 2.4 displays the distribution of reading items across booklets according to the three sub-domains.

TABLE 2.4 Number of reading items by sub-domain and by booklet

Sub-domain	Booklet 1	Booklet 2	Booklet 3	Booklet 4
Comprehension	4	3	4	3
Interpretation	3	3	2	3
Response to text	1	2	2	2
Total number of items	8	8	8	8

Linking the 2007, 2010, and 2013 reading assessments

PCAP aims to determine whether the performance of students changes over time. However, this type of comparison presents significant challenges. The major focus of PCAP rotates among the three administrations in a cycle. Because of this rotation of major/minor test focus, the tests in reading are not identical in successive assessments. Reading was the major domain in 2007 and included a large number of items, which enabled broad coverage of the sub-domains delineated in the PCAP Reading Assessment Framework. In 2013, as was the case in 2010,

reading is a minor domain with a limited number of items (approximately 20 per cent). Although items were selected from each sub-domain and with a range of difficulties (*p*-value range: 0.37 to 0.82; correlation range: 0.25 to 0.63) and performance levels (1 to 3),[6] the use of a smaller set of items means that the framework coverage is less representative. To facilitate comparison between assessments, the 2013 reading test was constructed from a subset of the 2007 items. These items, known as "anchor items," are used to link the 2007, 2010, and 2013 reading assessments and are used to report changes in reading achievement over time. These items, of course, were also a subset of the 2010 reading test, although there were five fewer items. Common items appeared between pairs of booklets to help judge the equivalency of the booklets for the reading domain.

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 caused variation in the 2007 means reported for reading between the two reports published in 2007 and 2010.

PCAP Mathematics Assessment Framework

The mathematics framework statement for PCAP 2013 was not altered from what was used to define mathematics performance in the 2010 assessment, where mathematics was the major domain.[7] This enables comparisons over time between the two cohorts.

Describing the domain

For this assessment, mathematics is broadly defined as a conceptual tool that students can use to increase their capacity to calculate, describe, and solve problems. The domain is divided into four strands or sub-domains and five processes. The PCAP assessment focuses on curricular outcomes that are common to all participating Canadian jurisdictions at the Grade 8/Secondary II level.

Regardless of the terms used to define mathematics, curricula across Canada are structured to enable a student "to use mathematics in his or her personal life, in the workplace, and in further study. All students deserve an opportunity to understand the power and beauty of mathematics. Students need to learn a new set of mathematics basics that enable them to compute fluently and to solve problems creatively and resourcefully" (NCTM, 2000, p. 1).

[6] Refer to the PCAP 2007 public report for the performance-level descriptors for reading at http://www.cmec.ca/Publications/Lists/Publications/Attachments/124/PCAP2007-Report.en.pdf

[7] See the assessment framework at http://www.cmec.ca/PCAP_Mathematics_Assessment_Framework

Organization of the domain

The mathematics component in PCAP 2013 is aligned with the jurisdictions' own curricula. The overriding principle of the assessment is that the application of mathematics is an integrated act in which the skills and concepts of various content areas are inherently linked.

The PCAP mathematics sub-domains are:

- numbers and operations (properties, equivalent representations, and magnitude);
- geometry and measurement (properties of 2-D figures and 3-D shapes, relative position, transformations, and measurement);
- patterns and relationships (patterns and algebraic expressions, linear relations, and equations); and
- data management and probability (data collection and analysis, experimental and theoretical probability).

Table 2.5 displays the distribution of mathematics items across booklets according to the four sub-domains.

TABLE 2.5 Number of mathematics items by sub-domain and by booklet

Sub-domain	Booklet 1	Booklet 2	Booklet 3	Booklet 4
Numbers and operations	4	0	0	4
Geometry and measurement	1	5	3	1
Patterns and relationships	0	2	4	3
Data management and probability	2	1	2	1
Total number of items	7	8	9	9

Mathematics curricula within the various jurisdictions in Canada include a number of mathematical processes deemed essential to the effective study of the subject. The processes reflect the means by which students acquire and apply mathematical knowledge and skills and are not intended to be separated from the knowledge and skills acquired through the curriculum content. These five processes are:

- problem solving
- communication
- representation
- reasoning and proof
- connections.

The sub-domains are traditional groupings of skills and knowledge, while the processes are used within all sub-domains.

For the PCAP 2013 mathematics component, test designers aimed to ensure that the contexts of the various scenarios were drawn from situations that were relevant, appropriate, and sensible for Canadian Grade 8/Secondary II students.

Linking of the 2010 and 2013 mathematics assessments

PCAP aims to determine whether the performance of cohorts of students changes over time, but because of this rotation of major/minor test focus, the tests in mathematics are not identical in successive assessments. Mathematics was the major domain in 2010 and included a large number of items, which enabled broad coverage of the sub-domains and processes delineated in the PCAP Mathematics Assessment Framework. In 2013, mathematics was a minor domain with a limited number of items (approximately 20 per cent) in this domain. Although items were selected from each sub-domain and with a range of difficulties (*p*-value range: 0.20 to 0.86; correlation range: 0.23 to 0.67) and performance levels (1 to 4)[8] the use of a smaller set of items meant that the framework coverage is less representative. To facilitate the comparison between tests in different years, the 2013 mathematics test was constructed from a subset of the 2010 items. These "anchor items" are used to link the 2010 and the 2013 mathematics assessments and to report changes over time in mathematics achievement.

Working groups

Working groups consisted of experts in reading, mathematics, and science. They came from various jurisdictions, 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.

Item Development

Documentation to guide all stages in the item-development process was prepared for the meeting of test developers in Toronto in September 2011. In preparation for the meeting, the science framework that had been developed at the beginning of the PCAP program in 2008 was revised to reflect changes in the science program of studies in jurisdictions across Canada.

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

The orientation included an overview of the science 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 in that small groups worked to develop a unit that contained a series of questions around a stimulus that had a good fit to both a sub-domain and a competency in the science framework. The units were presented to the large

[8] Refer to the PCAP 2010 public report for the performance-level descriptors for mathematics at http://www.cmec.ca/Publications/Lists/Publications/Attachments/274/pcap2010.pdf

group for discussion of 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. 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. 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 sub-domains and competencies.

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

After the item-development session, a number of sub-domain topics were not yet represented. Two independent contractors with extensive experience in item development were approached to write items on specific topics following the same development guidelines that the item-development working group used.

Because not all jurisdictions were able to participate in the item-development process, CMEC invited jurisdictions to donate units of items that could be considered for the science assessment. Items had to fit the science framework on topics that were common to Grade 8/Secondary II students in Canada. These items could be from jurisdiction or local assessments of grades adjacent to Grade 8/Secondary II, as long as they were accompanied by permission to adapt them to suit the PCAP assessment and were free of copyright restrictions. Items were subsequently donated by British Columbia and Alberta.

A research study was proposed to attempt to provide longitudinal data in science by devising a link between the PISA and PCAP assessments. Several PISA public domain units that focused on scientific literacy and that fit the PCAP science framework were selected to be included in the field test.

Item translation and review

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

An item review working group was held in Toronto in February 2012 made up of representatives from Manitoba, Quebec, Prince Edward Island, and Newfoundland and Labrador. The meeting's goal was to review the science items for content, vocabulary, translation, program of studies fit, and freedom from bias, and to verify the classification of the items for the competencies and sub-domains delineated in the science framework. The

committee 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 the committee members. As a result of this committee's 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, the science framework was again reviewed and revised to better reflect the common elements of the program of study documents from the jurisdictions. In some cases, changes to curricular documents over the course of the item-development process resulted in the removal of question units focused on topics that were no longer common to all Canadian Grade 8/Secondary II programs of study. Other 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.

Editing and verification of the assessment items

Before including items in an assessment, whether for the field trial or the main PCAP administration, it was important that these items be edited from various perspectives by a group of experts. They had to take as much care as possible to ensure that items were sound and would provide an accurate assessment of the skills of Grade 8/Secondary II students across the country. When editing items, the groups of experts took various steps: translating and comparing items in English and French, editing for language and style, scientific 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 Science, context or scenario 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.

Readability of PCAP contexts and items

The jurisdictional coordinators' meeting in fall 2012 raised concern that students writing PCAP in French, especially students who were members of minority French populations, might have a disadvantage in the PCAP science assessment based on language. They recommended that the science contexts and items be further scrutinized for the possibility of a bias based on language that favoured native English-speaking students.

Determining readability

There are a number of algorithm-based models that can be used to analyze a text's readability. The Flesh-Kincaid model[9] is based on a formula involving the number of words, number of sentences, and number of syllables. The Kandel and Moles formula[10] is a modified version that is adapted for French text. Both of these models calculate the text's ease of readability and assign it a value between 0 and 100, with higher numbers related to easier texts. The Flesh-Kincaid model also calculates a grade-level formula that translates the readability score to a US grade level.

Teachers and curriculum specialists were consulted regarding their use of such models when writing documents and tests for students. Generally, such tests are used as guidelines for comparative purposes but are treated with caution. According to Benjamin,

> while past researchers designed hundreds of formulas to estimate the difficulty of texts for readers, controversy has surrounded their use for decades, with criticism stemming largely from their application in creating new tests as well as in their utilization of surface-level indicators as proxies for complex cognitive processes that take place when reading a text. (2012, p. 1)

In other words, reading is a complex cognitive process so readability is more complex than an algorithm of the number of words, syllables, and sentences. Other important factors may include word familiarity, repeated words, and common terms used in the target subject. Thus, professional judgment was more important than the readability tests, although the tests can be used for guidance rather than as the sole criteria.

For the PCAP science items, the Flesch-Kincaid grade-level formula, a function in MS Word, was used to determine grade level for each unit. The range in reading grade level was determined to be between 3.5 and 8.9 with the majority of the units at the US Grade 8 reading level. A Web-based program, the Readability Index Calculator, was used to calculate readability because it allowed both English and French readability to be determined. For English, the Flesh-Kincaid formula was used while the French items used the Kandel and Moles formula.

[9] Formula proposed by Klare (1988).

[10] Formula proposed for French by Kandel and Moles (1958)

Readability test results

The Readability Index Calculator tool reported higher grade levels and lower readability ease than the MS Word tool even though both were using the same algorithmic model. This supports educators' concerns regarding the inconsistency of such algorithm-based models that oversimplify the process of reading. The Readability Index Calculator does compare reading ease for the two languages, which provides valuable information. For example, the comparison revealed that the concern about students writing in French having more difficult text was unwarranted. Only one unit was easier for the students writing in English and a second unit was at the same reading level in both languages. All other units were easier to read for students writing in French. Thus, based on algorithm-based readability tests, the text in PCAP 2013 Science items had higher readability ease for students writing in French than in English. The results of the readability analysis are reported in Table 2.6.

TABLE 2.6 Comparison of readability of PCAP science contexts and items

	MS Word	Readability Index Calculator (Web-based)	
Scenario name	US Grade Level (Flesch-Kincaid)	Readability in English (Flesch-Kincaid)	Readability in French (Kandel & Moles)
Air Pollutants	8.6	42	55
Alternative Energy	8.1	30	47
Animal Tracks	6.2	43	64
Ant Farm	6.9	29	48
Baseball	7.5	49	65
Beetles	6.2	61	59
Buying Locally	8.8	42	53
Climate Change	8.3	37	60
Crime Scene Investigation	6.7	40	79
Deforestation	8.5	22	38
ELA	8.4	37	64
El Niño	8.7	36	69
Expansion Joints	7.3	41	60
Extinction	9.8	29	62
Giraffes	7.5	39	56
Global Warming	8.8	32	63
Hot Work	7.2	58	67
Ice Sculpture	8.2	50	61
Icebergs	7.0	68	79
Impact of Fertilizers	5.2	39	60
Importance of Water	6.5	32	54
Skateboards	8.9	30	47
Solutions	8.3	40	50

States of Matter	7.8	34	54
Stream Bed	4.6	59	71
Summer Cabin	5.8	51	63
Tooth Decay	7.2	34	65
Tracks in the Snow	3.5	41	41
Trees and Forests	8.1	35	61
Water Impact	8.8	22	62
Weather & Erosion	6.8	50	70

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 point 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.

Scientific editing

Scientific editing checks and validates the correct answers, calculations, data, etc. The four versions of the test contained several 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 constructed-response items by experts in science. The coding guides used for reading and mathematics remained unchanged from the previous administration in which each subject was the major domain (2007 for reading and 2010 for mathematics) 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.

Item approval by the jurisdictions

Before including items in the field test, the jurisdictions had to approve of the items selected. CMEC produced three field test booklets, in English and French, and then sent these to the jurisdictions for their review. CMEC obtained approval from each jurisdiction to include the scenarios and items in the field test.

Chapter 3. Development of the Contextual Questionnaires

Initial questionnaire framework and guiding principles

To prepare for PCAP's initial design in 2007, the working group first reviewed sample questionnaire designs in three large-scale assessment programs: SAIP, TIMSS, and PISA. The group felt that any questionnaires designed for PCAP should be shorter and have a more explicit focus than those used in SAIP and PISA. In particular, they argued that student time was at a premium and that student questionnaires should be significantly streamlined. They also agreed that, to maximize research value, the questionnaires should be designed around some specific research focus rather than taking the broad approach of earlier questionnaires.

They therefore adopted the following principles when designing the questionnaires:

1. Include in the questionnaires some core descriptive data useful for both policy and research (e.g., student SES, school demographics, and teacher qualifications).
2. Other than core data, do not duplicate PISA.
3. Attempt to probe fewer areas in greater depth.
4. Identify policy-relevant issues.
5. Exclude areas that SAIP and PISA found non-productive.
6. Focus on the major domain in developing questions around teaching and learning strategies and behaviours.
7. Identify a limited number of areas that support the directions identified by the Pan-Canadian Education Research Agenda (PCERA).

The working group examined the limitation imposed by the short-term cross-sectional nature of the data on teaching and learning and agreed that an attempt might be made to ask questions designed to get at students' longer-term schooling experience.

There was no clear sense that any of the existing frameworks were inherently better than others. In the same way that the PCAP assessment is neither explicitly curriculum-based nor literacy-based, a more eclectic approach to questionnaires is required, based on identified research priorities and the need to link the questionnaires to the major domain.

A working group was convened in Toronto in January 2012 to develop questionnaires. The committee was made up of three external experts representing academic institutions and an organization promoting scientific literacy as well as four jurisdiction representatives from Manitoba, New Brunswick–anglophone, New Brunswick–francophone, and Newfoundland and Labrador. The members had strong expertise in science content, education research, statistics, and questionnaire item development.

The meeting's goal was to develop three concise questionnaires that focused on issues related to learning and teaching science, which was the major domain, and that could provide important contextual information for the jurisdictions. In preparation for developing the items, the committee reviewed science questionnaire items from a number of sources: previous administrations of PCAP (focused on reading and mathematics), SAIP, PISA, and from the

academic literature. The working group started with the concept that the PCAP questionnaires had to be shorter and more targeted, and include questions that were likely to produce interesting data. Decisions about retaining items for background and demographic information that were used in previous PCAP administrations were made based on the data from questionnaire analysis. By removing items that were used during the previous two administrations but that did not yield information, the group created more concise materials. The science-focused questionnaires developed by the working group were translated and copy-edited by CMEC and sent to the jurisdictional coordinators for review and further revision.

There were three questionnaires included in the PCAP 2013 assessment: one for participating students, one for their Grade 8/Secondary II science teachers, and one for school principals. The overarching structure of the three questionnaires was derived from the Wang-Haertel-Walberg synthesis of research on factors associated with school learning (Wang, Haertel & Walberg, 1990, 1993, 1994).These questionnaires also focused on the particular need to capture factors associated with science 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, science, 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 science.

Core questions

The core section included a limited number of questions for descriptive purposes and for comparison or control variables in research models. Some of the topics addressed in the Student Questionnaire included student gender, Aboriginal status, home background, SES, immigration status, home language, and language of instruction. The Teacher Questionnaire included teacher demographics, qualifications and assignments, and professional development, while the School Questionnaire included school demographics and governance, community context, and composition of the student body. PCAP 2007's questions on home language were found to be insufficient to pursue that area at the level of detail required for a special report on the achievement of majority and minority official-language groups, so this area was considerably expanded for PCAP 2010. However, when few trends were found, it was reduced again for 2013.

Gender differences

Differences in reading achievement favouring girls have been a consistent feature of large-scale assessments, both nationally and internationally. Differences in science and mathematics achievement tend to favour boys but are much smaller than the reading differences. 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 science-related behaviours or interests outside of school.

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

Time allocation and use

The topic of time has been a major feature in some other assessments. There is also a strong theoretical and empirical basis for time as a contributor to achievement. PCAP is trying to find ways to enhance its ability to measure time allocations and time loss by omitting previous variables that have little variance (e.g., length of school year) and by asking more specific questions about time management and student engagement in school. These include:

- time on specific subject areas
- length of class periods
- homework assignment and completion
- time lost (days, class periods, within class sessions)
- out-of-school time relevant to learning
- absenteeism
- exam times.

Science teacher efficacy and beliefs

Teacher efficacy is defined as "teachers' confidence in their ability to promote students' learning (Hoy, 2000, p. 1). The survey developed by Riggs and Enochs (1990) was included in the Teacher Questionnaire to explore the influence of teacher efficacy on student achievement in science.

Assessment

Many jurisdictions have responded to concerns about the performance of students and schools by implementing jurisdictional assessment programs. These take different forms and are of different degrees of maturity in different jurisdictions. Assuming that the underlying goal of this policy direction is to improve and not merely to describe achievement or entrench current levels, there is strong reason to examine assessment practices in the jurisdictions, and particularly how jurisdictional assessments are used. The intent here was to expand the scope of questions about assessment. Some areas for question development were: assessment practices, teacher knowledge of assessment principles; and measuring different levels of thinking (e.g., knowledge of facts, ability to apply knowledge, design scientific investigations, or evaluate information).

Accommodations (adaptations) and modifications

Previous PCAP questionnaires had a set of questions addressing some of the research and policy issues surrounding how to adapt instruction and assessment to meet all students' needs in classrooms. The broad policy context for this area is the strong movement in most jurisdictions toward including students with special requirements (e.g., physical, emotional, or intellectual challenges) in regular classes. Questions in the recent PCAP focused on differentiating instruction to accommodate various learning styles and providing support or assistance to students within classrooms.

Attitudes and motivations

This area is examined in some detail in PISA. Questions and constructs in this area are consistently found to be related to achievement. However, it has been considerably streamlined in PCAP. Because this area can be adequately researched using PISA, there is no need to duplicate PISA's approach. The basic idea here is that PCAP should include only the minimal number of items needed to permit use of attitudes and motivations as control variables in research on teaching and learning strategies. Items were developed on attitudes (general and subject-specific), interest, and self-concept.

Student learning strategies

The study of student learning strategies is considered one of PCAP's core elements. The questions in this key area dealt with student cognitive and meta-cognitive strategies in science, that is, the science strategies that students use when confronting different tasks and at different levels of difficulty.

Teaching strategies

Both the SAIP and PISA questionnaires included lengthy lists of questions about teaching strategies to which students (and teachers in SAIP) were asked to respond. These included generic questions about disciplinary climate, use of time, and student-teacher interactions, as well as more subject-specific questions. Typically, they were about the student's or teacher's experience in a particular class in the year of the survey. Because of this narrow scope, it seems likely that this results in systematic underestimation of the effects of teaching. Rather than simply duplicating the kinds of items found on the SAIP and PISA questionnaires, an attempt was made in designing the PCAP questionnaires to "reach back" to capture the student's longer-term classroom experience. While this will likely be difficult to do, it can, if successful, contribute to our understanding of students' broader school experience and how this relates to their achievement.

Thus, another small set of questions dealt with teachers' and students' perceptions that are purported to contribute to science achievement. PCAP gathered additional information about teaching strategies by asking students about their attendance at school and about their teacher's classroom practices (subject-specific). Questions in this section include:

- teacher perceptions of what contributes to science achievement

- student perceptions of their earlier school experiences
- experiences with science, and school questions on overall instructional philosophy and approach to science learning.

Opportunity to learn

Since opportunity to learn has often been considered one of the better predictors of achievement, a small set of questions aimed to determine:

- students' individual histories of being taught science
- parental activities related to opportunities to learn

One interesting feature of the PCAP 2013 Grade 8/Secondary II assessment results was that linking student performance to the three questionnaires permitted direct association of the output data (performance results) to the contextual elements for which information was gathered.

Item types

In the Student Questionnaire, School Questionnaire, and Teacher Questionnaire, most questions presented a range of answers—participants could generally check only one. The questionnaires also included several opinion questions to measure attitudes and reactions. These questions were based on a Likert scale to quantify attitudes. A Likert scale is an ordinal scale on which the answers to a question are ranked in order. A series of statements was presented to participants for which they had to indicate their level of agreement with each statement (e.g., "totally disagree" or "totally agree"). The questionnaires also included a few items in which participants had to write an answer, such as "number of hours" or "number of days on which…"

Contextual questionnaires

Student Questionnaire

Once students finished the assessment, they had 30 minutes to answer the questions in the Student Questionnaire. Most of these questions were related to science since that was the main PCAP assessment domain. Approximately 32,000 students responded to the questionnaire.

There were four sections in the student questionnaire on the following topics.

1. The student's personal information, either about the student's parents or guardians or about himself or herself. These questions gathered demographic data about the student (e.g., gender, socioeconomic status, immigrant status).
2. Students' attitudes and motivation, since these are generally related to performance. These questions looked at attitudes toward school and how students valued science both in their personal lives and in society.
3. The students' experiences learning science both in class and when they were younger.

4. The breakdown of a student's use of time. For example, some questions asked students how much time they spent on homework or other activities, and about absenteeism.

The data compiled from the Student Questionnaire will support a comparison and draw links between the variables studied as well as the student's performance.

Teacher Questionnaire

It is equally relevant to gather information from teachers on the variables studied in the assessment. The Teacher Questionnaire was filled out by the science teachers of Grade 8/Secondary II students selected to take part in the assessment.

The Teacher Questionnaire contained six sections on the following topics.

1. Personal information about the teachers, such as gender, education, and experience.
2. Professional development (PD), including questions about the number of PD days and both general and science-specific PD opportunities.
3. Time management, including how often they assigned homework, and the reasons that learning time was lost in classrooms.
4. Assessment practices (e.g., types of assessment, how marks were assigned, assessing types of thinking, and how the various needs of students were accommodated).
5. Teaching strategies, for example, how the classroom was organized for teaching, differentiation methods, and the frequency and use of science-related strategies and activities.
6. Science-teaching efficacy and beliefs and perceived challenges to teaching. The questions covered the teacher's attitudes toward science, including reasons that students do more or less well in science, teacher's self-confidence, self-evaluation, and challenges.

School Questionnaire

The school's principal in schools selected to participate in PCAP were asked to complete the School Questionnaire. This provided information at the school level in relation to the PCAP 2013 science assessment and allowed for analysis and links with the students who completed the assessment.

The School Questionnaire was divided into five sections on the following topics.

1. General information. For example, questions covered the number of students enrolled in the school, the grades taught there, the percentage of Aboriginal students, and the size of the community where the school was located.
2. Time management. Questions covered such issues as the time allocated to science instruction and absenteeism.
3. Assessment. Principals where asked about accountability practices and challenges to teaching science in their schools.
4. The instructional climate. Questions about instructional emphasis and the promotion of science. Respondents indicated their level of agreement in relation to the science

teaching atmosphere in their school and indicated the frequency of different events (e.g., professional development, parent nights, science fairs, and recognizing student achievement).

5. The context for instruction, including differentiated instruction and instructional challenges.

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 2013, the third Pan-Canadian Assessment Program (PCAP) was administered. It assessed three domains: science, reading, and mathematics, with science being the primary domain. Four assessment booklets were used in which all three domains were assessed, with the majority of the items focusing on science. 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 plan

Between 1993 and 2004, the Council of Ministers of Education, Canada (CMEC), became involved with pan-Canadian assessments through the School Achievement Indicators Program (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. In 2010 and 2013, 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 and to compile mean results.

The sampling plan had to be adapted to this context. As was the case for the SAIP assessments, a two-step procedure was followed. First, participating schools were selected, and second, a Grade 8/Secondary II class was chosen in the selected schools. Given the size of the populations being assessed, a census could be 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.

Because the statistics produced for students in a sample had to be generalizable, each sample had to meet certain criteria.[11] 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. Table 4.1 provides statistics on the various Canadian populations targeted by PCAP. These statistics are derived from data that the jurisdictions supplied to CMEC for the May 2012 field test of the assessment.

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

TABLE 4.1 Number of students in Grade 8/Secondary II, by population[12]

Population	Number of schools	Number of students
British Columbia — anglophone	563	46,632
British Columbia — francophone	13	263
Alberta — anglophone	823	39,820
Alberta — francophone	23	274
Saskatchewan — anglophone	524	12,498
Saskatchewan — francophone	49	100
Manitoba — anglophone	423	14,070
Manitoba — francophone	21	381
Ontario — anglophone	2,756	139,680
Ontario — francophone	173	6,076
Quebec — anglophone	124	8,472
Quebec — francophone	562	76,963
New Brunswick — anglophone	88	5,727
New Brunswick — francophone	61	2,249
Nova Scotia — anglophone	148	9,424
Nova Scotia — francophone	11	368
Prince Edward Island	29	1,487
Newfoundland and Labrador	154	5,441
Total	**6,545**	**369,925**

The PCAP assessment plan assumed that the test would be administered in an optimal number of schools. The following are some parameters relating to the sampling plan for the main study, which took place in May 2013.

[12] The statistics provided by the jurisdictions for field testing PCAP made it possible to construct this table.

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 were selected to guarantee a margin of error of no more than 3 per cent, with a confidence level of 95 per cent. Table 4.2 gives the formula used to determine the size of a sample in relation to the calculation of frequency distributions.

TABLE 4.2 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

Use of multiple assessment booklets

In PCAP 2013, four assessment booklets were used. They were designed according to the domains assessed, number of items per domain, and the overall difficulty per booklet.

Table 4.3 Structure of the assessment booklets

Booklet	Science	Reading	Mathematics	Total Items
1	24	7	7	38
2	22	8	9	39
3	24	9	8	41
4	24	9	8	41

When the data were processed, the responses from each booklet were compiled. It was necessary to merge the results obtained to create a database containing results for all the students assessed. Several weighting values were calculated so that the structure of the tests employed could be taken into account.

The use of several booklets made it possible to gather more information concerning different sub-domains. Since the statistics produced for these sub-domains take into account students' performance in relation to mean results, the associated margins of error were acceptable. This is not always the case for frequencies that concern the overall performance of sub-groups of students for each domain assessed. The proportions of students associated with some groupings of performance levels will, in some cases, probably have margins of error greater than 3 per cent for a confidence level of 95 per cent. The compilations done for the complementary Student Questionnaire allowed analysts to obtain a margin of error lower than 3 per cent.

Several populations were assessed as part of this activity. Given their size, a number of possibilities had to be considered. For the selected schools (the first level of the sampling plan), this involved a fixed number of schools or a census of all the schools belonging to a target population.[13]

Existence of several sampling specifications

Table 4.4 shows the distribution of populations on the basis of whether they had sampling at the first level, that is, whether there was a selection (column 2) or not (column 3) of schools for the assessment. If there was a census of schools, there may also have been a census of students (column 4).

[13] "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 jurisdictions concerned. The "overall population" consists of all the schools that have Grade 8/Secondary II students.

TABLE 4.4 Distribution of populations, by first-level and second-level sampling

1	2	3	4	5
Population	**Sampling at the first level**	**Census of schools**	**Census of students**	**Number of students to be evaluated**
British Columbia — anglophone	X			3,300
British Columbia — francophone			X	259
Alberta — anglophone	X			3,300
Alberta — francophone			X	262
Saskatchewan — anglophone	X			3,300[14]
Saskatchewan — francophone			X	100
Manitoba — anglophone	X			3,300
Manitoba — francophone			X	376
Ontario — anglophone	X			3,300
Ontario — francophone	X			2,000
Quebec — anglophone		X		2,000
Quebec — francophone	X			3,300
New Brunswick — anglophone		X		2,000
New Brunswick — francophone		X		1,000
Nova Scotia — anglophone		X		2,500
Nova Scotia — francophone			X	368
Prince Edward Island		X		800
Newfoundland and Labrador	X			1,500

[14] 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. Schools in anglophone school districts that offer both English- and French-language programming are known as dual schools.

Table 4.5 present the parameters used to determine the size of the various samples, given the types of populations.

TABLE 4.5 Sample size parameters

Type of population*	Sampling parameters
Populations sampled at two levels (column 2)	Approximately 150 schools; one Grade 8/Secondary II class; one quarter of students will be assessed using one of the four booklets[15]
Census of schools (column 3)	One Grade 8/Secondary II class for every school on the list; one quarter of each class will be assessed using one of the four booklets
Census of students (column 4)	All the students on the list; one quarter will be assessed using one of the four booklets

* Note: column numbers refer to Table 4.4

Choice of schools

The CMEC data centre chose the schools that participated in PCAP. This selection was made by applying the same rules to each one, using information provided by the jurisdictions.

Databases on the schools

To carry out the sampling work, CMEC needed a database for each population assessed. Each jurisdiction had to use the same file prepared by CMEC to draw up the list of schools and prepare other necessary information.

Selection of schools

CMEC decided to centralize some operations related to the various samples of schools selected for this assessment. This strategy ensured greater uniformity in the techniques employed to perform the required operations. This also made it possible to take the sampling process into account.

For the larger populations, namely, those presented in column 2 of Table 4.4, students were selected from about 150 schools. For these populations, the sampling plan was at two levels (schools and students). The schools were chosen randomly and their size was taken into account. At the time the schools were selected, the available strata were considered. Any given school was selected only once.

When the statistics were compiled, weights were assigned to each student, using an identical technique for all populations. This same technique was also used to calculate margins of error.

[15] Several classes in the same school could be selected on an exceptional basis.

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 jurisdictions 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 were 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 were 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 jurisdictions 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 jurisdiction other than the province or territory 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.

Selection 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. CMEC 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 jurisdictional coordinator submitted a list of all eligible schools with Grade 8/Secondary II students that were under provincial/territorial jurisdiction.
2. CMEC selected schools to participate in PCAP and sent the *List of Schools* to provincial and territorial 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 and territorial coordinators. It was possible that, in some cases, more than one class was chosen in the same school. After consultation with schools, jurisdictional 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. Jurisdictional 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, jurisdictions sent CMEC a list of students who participated in PCAP 2013. The same lists prepared for step 5 were used, with reasons given for any student's non-participation in the assessment.

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 and territorial coordinators. When it comes to students to be assessed, CMEC selects the class(es) in each school that is part of the chosen samples.

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-selection working group—field study

The item-selection working group, with representatives from four jurisdictions, met to review all science items for content, vocabulary, and translation, and verify the items' classification for the test's sub-domains and competencies. 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 science sub-domains and competencies with a range of difficulties.

Assessment booklets

Three booklets were compiled for the field test. Each booklet followed the specifications outlined in the *PCAP Science Assessment Framework* and contained about 40 science 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.

Item-scoring session

The scoring session took place over four days in Gatineau from July 10 to 13, 2012. 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 14 scorers, half of whom were assigned to the English table and the other half to the French table.

The scoring 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 2.

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 jurisdictions 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 science 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, sub-domain, and competency 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 jurisdictions, met to review and select scenarios 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 both genders. The item-selection process also took into account comments from scorers 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: science, reading, and mathematics. A very small number of science items were retained from previous PCAP administrations but were not intended to be used as anchors. All reading and mathematics items were anchor items and as such were identical to those used in the cycle in which these subjects were the primary focus of the assessment: PCAP 2007 for reading and PCAP 2010 for mathematics. Because only a small subset of items was required, the working group took care to represent each sub-domain 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 Science 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 29 and May 24, 2013, with the primary domain being science. (The minor domains were reading and mathematics.) More than 32,000 students selected at random from close to 1,600 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 science questions, and one cluster each of reading and mathematics items. In order to assess the equivalency of each booklet, a sub-set 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 jurisdictional coordinators reviewed them so that comments could be incorporated as required. The materials sent to the jurisdictions 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 jurisdiction.

Printing the assessment booklets

After all jurisdictions 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 sent to the jurisdictional coordinators for distribution to the schools.

Checking documents

Each jurisdictional 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 jurisdictional coordinator directly.

Letter sent to parents/guardians of students

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

Administration procedures

Each school selected had to appoint a school coordinator to administer PCAP in that school. The assessment then had to be administered according to the procedures that CMEC established to ensure that PCAP was administered uniformly in all the selected schools. Before proceeding with the assessment, the school coordinator had to become familiar with the administrative documents, in particular the *Handbook for Schools*, which outlined the test's administrative procedures. If the school coordinator had any questions related to the assessment, s/he had to communicate with the jurisdictional 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, s/he was 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 him or her 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 he or she had participated in, had been exempted from, or was 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 jurisdictional 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 *Jurisdictional 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 jurisdictional 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 online versions. In some jurisdictions, 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 science 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 online) 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 science, reading, or mathematics 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 and territory. 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:

1. 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.
2. a break, or multiple breaks, as long as students are supervised during the breaks
3. an alternative setting
4. use of Braille, large-print, coloured paper
5. use of a scribe (writing verbatim: must write what student says without editing)
6. verbatim reading of instructions only, for all domains
7. 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 I-2 in Appendix I of the public report (O'Grady & Houme, 2014)[16].

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:

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

[16] http://cmec.ca/Publications/Lists/Publications/Attachments/337/PCAP-2013-Public-Report-EN.pdf

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 the jurisdictional coordinator 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 Moncton, New Brunswick from July 8 to 19, 2013. All the science, reading, and mathematics items were scored by teachers in the relevant domains. Science-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 reading items were scored during the second week. In all, there were 120 scorers, both anglophone and francophone.

In total, over 32,000 assessment booklets were scored, with approximately 24,000 in English and 8,000 in French.

Bundling booklets

All the assessment booklets were bundled before scoring the items. A bundle contained 10 assessment booklets from various jurisdictions. 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 jurisdiction the assessment booklets came from, to avoid any bias in scoring the items toward one jurisdiction 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, scoring 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 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 four to six 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 science, reading, or mathematics 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 jurisdictions 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. During the scoring session, if score 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 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.

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 results showed us that most scorers obtained a more-than-acceptable level of agreement with the experts. The reliability review results were satisfactory for all three subjects assessed. The percentage agreement by language and overall for each scoring group and domain are presented in the four tables below.

TABLE 6.1 Reliability review results for science scoring group 1

	Cluster 1		Cluster 2		Cluster 3			Cluster 4				% agreement
	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3	Test 4	
English	90	90	88	99	83	93	-	85	94	98	97	92
French	92	85	93	96	95	94	95	82	80	98	93	91
Overall	91	88	91	98	89	94	95	84	87	98	95	91

TABLE 6.2 Reliability review results for science scoring group 2

	Cluster 1		Cluster 2				Cluster 3				Cluster 4				% agreement
	Test 1	Test 2	Test 1	Test 2	Test 3	Test 4	Test 1	Test 2	Test 3	Test 4	Test 1	Test 2	Test 3	Test 4	
English	81	88	83	70	92	98	74	75	95	-	96	95	96	-	87
French	99	98	91	94	91	98	82	87	96	88	86	93	88	96	92
Overall	90	93	87	82	92	98	78	81	96	88	91	94	92	96	89

TABLE 6.3 Reliability review results for reading

	Cluster 1				Cluster 2			% agreement
	Test 1	Test 2	Test 3	Test 4	Test 1	Test 2	Test 3	
English	72	85	89	73	69	74	78	77
French	83	84	79	95	81	89	88	86
Overall	78	84	84	84	75	82	83	81

TABLE 6.4 Reliability review results for mathematics

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

Inter-rater reliability (double scoring)

Double scoring was a quality control measure in scoring assessment booklets for science, reading, and mathematics in which about 2,400 booklets (300 of each booklet in English and in French) were scored a second time by another scorer. As Table 6.5 shows, the overall consistency between scorers is 95 per cent in both science and mathematics and 80 per cent for reading.

TABLE 6.5 Overall agreement between coders for double scoring

Booklet	Domain and scoring group			
	Science Group 1	Science Group 2	Reading	Mathematics
1	99	98	92	99
2	95	95	81	97
3	95	94	71	93
4	90	92	77	90
Average	95	95	80	95

Trend reliability

Trend reliability was a quality control measure to estimate the degree of agreement between mathematics and reading scorers for the anchor items in PCAP 2013 and the same items in PCAP 2010. Four items for reading and nine items for mathematics were common between the two tests. About 2,400 booklets from 2010 (300 of each booklet in English and in French) 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 or equal numbers of each booklet to a maximum of 300. 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.

Multiple scoring

During the scoring of PCAP 2013 the two clusters of science items in each booklet were scored by different groups of scorers. As a check to ensure that the two groups' scoring of the booklets was equivalent, one booklet was scored by all scorers.

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 jurisdictional coordinators. Scorers provided feedback during the scoring session. The information collected in these reports is summarized the following sections.

Jurisdictional coordinator's report

Following the assessment's administration, the jurisdictional 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 jurisdictional reports was used to make any necessary changes to the administration process for future assessments. The jurisdictional coordinator's report included seven questions.

First, the jurisdictional coordinators had to summarize the methods the schools used to encourage students to participate seriously in the assessment. In most cases, the jurisdictions 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 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, it seems that these things may not have been made available to students in all schools. Schools were able to choose whether their class of French-immersion students wrote the assessment in English or in French but some schools indicated that such students writing in English were unfamiliar with the vocabulary in the test because their classroom instruction had been in French.

Jurisdictional 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. Jurisdictional coordinators were able to address some of these issues as they arose because they had extra booklets in both languages.

The jurisdictions' comments indicate that the majority of schools complied with the administration procedures. In most jurisdictions, 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 jurisdictional 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 2013 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 jurisdictional 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 jurisdictions and that it was important to effectively represent their school and province or territory. 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. Some schools indicated that the voice on the audio CD in English had an accent which was difficult for students to understand.

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:

- Some schools chose not to notify students and parents of the test prior to the test date.
- In one case, an invigilator did not have a return address and indicated that he discarded the test.

- 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 School 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.
- In a few schools, additional time was given beyond that allowed in the guidelines for assessment.
- 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 or three settings instead of being administered in one 90-minute setting with short breaks.
- In a few schools, the selected class was not a science class. Questions related to science teaching and learning in the teacher questionnaire were problematic for the teacher.

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.

Fifty-eight per cent of respondents scored science which was the major domain, 21 per cent scored reading, and 21 per cent scored mathematics, both of which were minor domains. Additionally, about 10 per cent of scorers scored in two domains.

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 science questions. They were also asked to compile a list of common misconceptions presented by the students in their

responses. The information was used in an issue of *Assessment Matters!* in which 14 science items from the PCAP 2013 Science Assessment were released with commentary on the student work and sample responses. [17]

[17] Available at http://cmec.ca/Publications/Lists/Publications/Attachments/339/AMatters_No8_PCAPItems_EN.pdf

Chapter 7. Setting a Performance Standard

Whenever tests' content and/or item types are modified significantly, standard setting should be 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 2013, science was the major domain in PCAP for the first 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 from November 25 to 28, 2013, in Toronto. The meetings were divided into three sessions: a one-day leaders' training session on November 25; two days of standard-setting sessions on November 26 and 27; and a one-day writing session to revise proficiency-level descriptors on November 26.

The standard setting aimed to articulate levels of performance on the PCAP science 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. Three cut scores were set to differentiate between four 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 four days, the cut scores recommended by the panelists were sent to the jurisdictional 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 jurisdictions were involved and that they had an opportunity to participate in setting cut scores. Each jurisdiction was invited to designate two representatives having some expertise in measurement and evaluation and in science content for Grade 8/Secondary II. The standard-setting committee consisted of 25 panelists. CMEC solicited committee members and standard-setting panelists through nominations by the jurisdictional coordinators. A key consideration of any such working group was that they 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 panel also consisted of 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 four 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 literature from international tests (e.g., TIMSS, PISA) informed how the preliminary PLDs were drafted. These were statements describing what students at the four performance levels knew and could do and were referred to by panelists 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 science framework and at this first step, panelists 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 coherent PLDs that the panelists agreed upon. Once finalized, the PLDs were ready for use at the standard-setting meeting.

Security of materials

Because standard setting uses operational material, 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 many 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 2013 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 science items that could be answered with a brief response that was scored correct or incorrect (code 1 or 0 respectively). ECR items were one- or two-point items that required a student's longer written response.

TABLE 7.1 Composition of the PCAP test booklets for science items

	Selected Response	Short Constructed Response (codes 0 or 1)	Extended Constructed Response (codes 0, 1, or 2)
Booklet 1	17	5	2
Booklet 2	18	4	2
Booklet 3	17	6	1
Booklet 4	17	5	2

With the Bookmark procedure, panelists 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 booklet, 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. Each page contained essential information about the item, including its position in the OIB, its position in the original booklet, and the achievement level (theta) required for a student to have a two-thirds chance or greater of answering correctly or obtaining that point.

CMEC used the Rasch model for item calibration and test construction. This model allowed for the calibration of all items and students on a common scale. This common calibration allowed for the calculation of the probability of a correct response to a given item by a given student from information about the student's achievement level (theta) and the item's difficulty level (p value).

Standard-setting procedure

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

The panel heard 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, and materials. 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 between the four levels.

Participants then took the rest of the morning and part of the afternoon to become familiar with the assessment instrument and the materials for the session. This step took some time,

but it was necessary for panel members to review the materials carefully to gain “fluency” with the assessment. Participants had discussions at their tables concerning the assessment items and item difficulty, and they were given an opportunity not only to review but also to answer items and score their answers, thereby gaining insight into performance-level descriptors.

The first bookmarking round took place before the end of the day, with participants reviewing each item in the OIB, discussing 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 item numbers denoting the three cut scores. The median of all responses defined the cut scores between levels 1 and 2, between levels 2 and 3, and between 3 and 4.

The second day began with a full group discussion on the first bookmarking round. CMEC staff posted the results (the numbers of all items bookmarked and the median used to establish the first, second, and third cut scores). Major variations were evident between participants’ responses, 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, there was an even greater difference between the lowest and highest cut score. However, the panel experienced less difficulty bookmarking the third cut score between levels 3 and 4. Many more participants had selected the same items, and there were fewer variations. The first round was a good exercise for participants and gave them an opportunity to share comments and opinions.

The second round was similar to the first, with participants placing bookmarks in the ordered item booklet to determine the three cut scores and providing a rationale for their choices. CMEC staff compiled results. Some participants decided to change their bookmarks, while others chose to leave them on the same item. There were fewer variations in responses for all three cut scores than in the first round. In the second round as well, participants appeared to experience less variation in setting the second cut-off point. For the second round (but not the first), participants were shown impact data, that is, the percentage of students performing at levels 1, 2, 3, and 4. Based on panel responses in the second round, approximately 2 per cent of students performed at level 1, 33 per cent at level 2, 53 per cent at level 3, and 11 per cent at level 4. Showing participants the impact data allowed them to check their choices against the outcomes and readjust their cut-off points accordingly. 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.[18] Average theta across the booklets was the statistic used to determine impact. Colour-coded bar graphs were prepared to illustrate the distribution of results for each cut level, for bookmark round 2 and round 3. 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. 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 science

Performance Level			
Level 1	**Level 2**	**Level 3**	**Level 4**
8%	44%	39%	8%

A questionnaire was distributed to participants at the end of the session to collect information, comments, and feedback on the standard-setting process and the method used, as well as on the assessment instrument itself. Most panel members reported that they had enjoyed the session, that they had been comfortable with the process, that the session had been an enriching experience, and that the bookmark method was a fair and easy-to-understand way to set cut scores. The majority of participants also appeared satisfied with the organization of the session and with the leaders and facilitators and they commented favourably on the assessment instrument. Most stated that the texts and questions were appropriate and that the science 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 four levels of performance. Using these items, they developed a description of the knowledge and skills that characterized achievement at each of the four performance levels.[19] Level 2 is considered the acceptable or "baseline proficiency," or the level at which students begin to demonstrate the competencies needed to participate in life situations related to science. Students achieving at level 1 are below what's expected of students in their grade.

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

[18] The theta statistic was adjusted for the 2/3 response probability as described earlier.
[19] These descriptions appear in the PCAP 2013 public report (O'Grady & Houme, 2014) available at http://cmec.ca/Publications/Lists/Publications/Attachments/337/PCAP-2013-Public-Report-EN.pdf

Chapter 8. Processing PCAP Data

Data processing is an important part of the project—it produces the assessment results! This is a fairly complex process because important 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. In Canada 32,604 Grade 8/Secondary II students wrote the assessment and responded to the Student Questionnaire, 1,594 science teachers of the participating students completed the Teacher Questionnaire, and 1,917 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 jurisdictions sent all the assessment booklets, answer sheets, and questionnaires to Toronto 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 Moncton, New Brunswick for the scoring session where the scorers corrected all constructed-response items in more than 32,000 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, the data entry company programed specific rules for each section and question response in the three questionnaires. After each 25th questionnaire was keyed, a quality control check was completed during which a supervisor confirmed the samples by rekeying the questionnaire. Any discrepancies were taken up with the operator prior to

rekeying and new batches. After all batches were keyed and processed, programming validated the output response using the rules established prior to data capture.

Data cleaning

When data were submitted by jurisdictions, the first step was to check the consistency of the database structure with the CMEC database. The data officer identified deleted variables, added variables, and variables for which the rules had been changed. All deviations were checked and verified with the jurisdictions. The data files were then sent to the CMEC data-processing centre for specific data cleaning or 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 jurisdiction, 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 deleted.
- A variable for an item that was deleted in the questionnaire because of a mistranslation was deleted.
- Duplicate data records were identified and only one record was kept based on decision rules established by the technical advisory committee. For example, for two files with the same identification code, the file that contained less information was 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 jurisdiction 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 by an external expert. The finalized databases were then sent to CMEC along with some preliminary analysis. 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 provided. This provided information about test items that appeared to have behaved in an unacceptable way and about any ambiguous data remaining in the questionnaires. With such information, the key was corrected 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. The data files were then weighted based on the population sizes outlined in the PCAP Sampling Framework.

Chapter 9. Analysis of Achievement Data

This chapter outlines the PCAP 2013 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 2013 reading and mathematics with PCAP 2007 and 2010
5. scoring and scaling PCAP 2013 performance data
6. standard error estimates
7. presenting the PCAP 2013 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) handling missing data. These steps were performed for each booklet with breakdowns by jurisdiction 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 jurisdiction. 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.[20]

Item recoding

The PCAP 2013 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.

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

The MC items in the English and French versions were separately recoded before the two data sets were merged. 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 2013 assessment: [21]

- 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.

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

Not 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. In addition, the examinee attempted no subsequent items in the remainder of the booklet.[22] 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 2013 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 s/he has 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.

[22] 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).

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 2013 items. For the major domain, which was science, 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 reading items were administered in both 2007 and 2010 and mathematics items were administered in 2010. These items were field tested when reading and mathematics were major domains. The statistics were reviewed in preparation for the selection of items to be included in PCAP 2013.

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 indexes (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 science domain. There were also eight tests, albeit smaller, for each minor domain. For these minor domains, the placement of the items in PCAP 2013 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 a 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 national-level estimates.

Reliability of the PCAP 2013 assessment

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

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[23]
- 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.

IRT analysis

The IRT analysis process involved: (1) assessing the dimensionality of PCAP 2013, (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 2013 invariance. The process ended with assessing the PCAP 2013 validity evidence. The validity assessment occurred through the differential item functioning (DIF) and the validity evidence was described under the DIF analysis.

Assessing the dimensionality of PCAP 2013

The PCAP 2013 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).

Two statistics programs, EQSIRT and 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.

In addition to FIML, the Fraser and McDonald (2003) Normal Ogive Harmonic Analysis Robust Method (NOHARM) was also used. This method has performed well in dimensionality recovery studies (de Ayala, 2009; see also De Champlain & Gessaroli, 1998; Finch & Habing, 2005; Knol & Berger, 1991). The three different types of software agreed on a dominant dimension underlying the PCAP 2013 major domain.[24]

Item calibration

Items from pairs of booklets were calibrated concurrently to link the booklets and to put both the items and the students on a common metric. This procedure makes it possible to estimate

[23] 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.

[24] It may be worth mentioning that SAS yielded similar results.

theta scores in a way that does not depend on the set of items to which the students responded.

Three separate item calibrations for the three domains were performed that involved estimating the reading, the mathematics, and the science item parameters separately. In addition, the reading and the mathematics calibration process used data from 2010 (when mathematics was the major domain) and the 2013 data simultaneously.[25]

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 competitive Birnmaum's three-parameter model (3PLM). The 2PLM was retained because the fit didn't improve significantly; 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.

Assessing the IRT models' fit

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 χ^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 χ^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-χ^2 statistics (IRTPRO does not produce and does not endorse producing the empirical item response curve). The S-χ^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 2013 items displayed different statistical proprieties (e.g., level of difficulty) for gender and language. 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"

[25] 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.

(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.

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. PCAP 2013 minor domains 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 2013. Therefore all reading and mathematics items were anchored. The design corresponded to the nonequivalent groups with anchor test (NEAT) design. However, because the change in the target population definition led to setting 2010 as baseline year for reading, the 2013 reading and mathematics assessments were both linked to the PCAP 2010 assessment. The linking was done by way of concurrent calibration. Because the parameters of two successive assessment items were estimated simultaneously with common items (e.g., 2010 and 2013 mathematics items), the common item parameters had the same estimates and were on the same metric (de Ayala, 2009; Kim & Kolen, 2006). The approach had the advantage of making maximum use of all the available data in estimating item parameters (Martin, 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 2013 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 (e.g., PCAP 2010 mathematics) achievement distributions estimated under the concurrent calibration to the same assessment distributions obtained when the item parameters were first estimated (e.g., the estimates used in 2010); and
- applying the linear transformation at stage II to the current assessment achievement distributions (e.g., PCAP 2013 mathematics).

The reading and the mathematics achievement score generation for PCAP 2013 (the theta scores and the scale scores) used the item parameters estimated at this stage.

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.

Achievement score generation and scaling

- 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 (science, reading, or mathematics). The estimation of the theta score used the Expected A Posteriori (EAP) method.
    - The science theta score generation used item parameters from the science item calibration.
    - For reading and mathematics, the concurrent calibration linking previous assessments and PCAP 2013 provided the item parameter estimates for theta score generation.
- The theta score was set on a scale with an unweighted national mean of 500 and a standard deviation of 100.
- The scores at stage II were weighted with the sampling weight, and the national mean reset to 500 with a standard deviation of 100.

The same three types of scores were also computed for all the students in each of the four science sub-domains and in the three targeted competencies.

Standard error estimates

The PCAP 2010 data analysis used a bootstrap approach in developing empirical standard error estimates for the national results and means by jurisdiction 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 changes 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.

Presentation of the PCAP 2013 achievement results

Summary score reports were developed at the national, jurisdictional, 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 and 2010 public reports.

Chapter 10. Analysis of Questionnaire Data

As in previous assessments, PCAP 2013 collected background data from students, teachers, and school principals (in the School Questionnaire). 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.

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

Frequency tables were produced for each item:

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

Item recoding

The PCAP 2013 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.

Missing data

Three types of missing data occurred in the PCAP 2013 questionnaires:

- missing responses because a student ran out of time to complete the questionnaire (not-reached); [26]
- 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.[27]

These types of missing data were coded 9. When it was possible, missing data were imput 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 jurisdiction and by language. They included frequency and percentage distributions for all items on non-continuous and Likert-type scales. The descriptive statistics also included the mean, the standard deviation, and the shape statistics (skewness).

Factor analysis

Factor analysis involved performing exploratory factorial analysis (EFA) of PCAP 2013.

Item analysis: Classical theory item analysis

A classical statistical analysis permitted the assessment of some measurement properties of PCAP 2013's background questionnaire items. It was conducted for each questionnaire and was reported by jurisdiction and by language. It primarily focused on the following items:

- mean and standard deviation
- item discrimination
- percentage of respondents with missing responses to the item
- internal consistency index for each scale of each questionnaire.

---

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

[27] The data from only 25 per cent of New Brunswick anglophone teachers and schools was captured because questionnaires completed on paper were returned several months after data capture was completed. Only questionnaire data submitted online by this population is included.

Group comparison analysis

The group comparison analysis involved:

- comparing achievement means or some student-related variables (e.g., student's attitude toward science, perceived efficacy, self-confidence) with regard to a given categorical variable (e.g., gender, school location — rural versus urban) from a questionnaire (Student, Teacher, or School Questionnaire).
- comparing performance level of achievement with regard to a given categorical variable from a questionnaire (Student, Teacher, or School Questionnaire).
- comparing achievement means with regard to a set of student-, teacher-, or school-related categorical background variables.
- comparing achievement means (other student variables such as attitude toward science, perceived efficacy, self-confidence, etc.) with regard to a set of teacher-related categorical variables.
- comparing achievement means (other student variables such as attitude toward science, perceived efficacy, self-confidence, etc.) with regard to a set of school-related categorical 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 derived variable (from a questionnaire), or a given scale background variable. Since the correlation coefficients were produced in matrix format, they also included the correlation among the derived variables and among the derived and other background variables.
- performing linear multiple regression analysis to predict achievement in science from a set of student-related variables, including categorical variables.
- performing linear multiple regression analysis to predict achievement at the class level, that is, the class mean achievement in science as well as achievement in reading and mathematics from a set of teacher- and school- related variables, including categorical variables.
- performing linear multiple regression analysis to predict achievement at the school level, that is, the schools mean achievement in science as well as achievement in reading and mathematics from a set of teacher- and school- related variables, including categorical 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 Databases

Description of the databases

All the PCAP 2013 databases are in English and French and are available to researchers. CMEC has several databases, 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 database containing all the student records merged with the questionnaire answers. This database can establish links between student performance and the contextual data. A description of the SPSS and Excel databases follows[28].

Student database

This database includes primarily the following data:

- general information about students (student, school, and teacher identification numbers; student participation code; booklet number; each student's province/territory and language);
- student and school statistical weights;
- responses to the Student Questionnaire;
- student factor scores;
- results of the achievement test (mathematics, science, and reading).

Teacher database

This database includes primarily:

- general information about teachers (teacher and school identification numbers; each teacher's province/territory and language);
- responses to the Teacher Questionnaire;
- teacher factor scores.

Definitive teacher population figures are not available because the teacher sample was based on the school and student samples. All teachers who taught science to students writing the PCAP test in a school were sampled; however, many teacher ID numbers are missing. In some jurisdictions, the questionnaire return rate was very low. Because intact classes were used, one teacher was sampled in most schools, with two or more teachers in a few schools.

School database

This database includes primarily:

- general information about schools (school identification number, each school's province/territory and language);
- school statistical weight;

[28] Codebooks which identify the variables and values in the databases are available upon request.

- principals' responses to the School Questionnaire;
- school factor scores.

Merged database — Student/teacher/school

This database includes all the information from the student, teacher, and school databases described earlier. This database will enable researchers to establish links between student performance and the contextual data.

Accessing the database 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 gave jurisdictions a simple way to compare their performance with that of the rest of Canada. PCAP data also provided information to jurisdictions about the performance of their own education systems.

CMEC is committed to encouraging policy-relevant 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 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 jurisdiction (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 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.

Terms and conditions

CMEC will maintain a registry of all requests for the use of PCAP data so that jurisdictions 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 databases:

- 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 jurisdictions that are named or identified in any research findings one month prior to the publication or release of the findings, so that the jurisdiction(s) 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.

Contact information

For more information about PCAP data, please contact pcapinfo@cmec.ca

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 (CMEC). (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 (CMEC). (2005). *The pan-Canadian assessment program: Literature review of science assessment and test design*. Toronto: Author (unpublished report).

Council of Ministers of Education, Canada (CMEC). (2011). *PCAP-2010: Pan-Canadian assessment program*. 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). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Psychological Measurement, 19*(4), 357–381.

Klare, G. R. (1988). The formative years. In: 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.

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.