Medical Council of Canada Qualifying Examination (MCCQE) Part I

2017 MCCQE Part I Annual Technical Report

MEDICAL COUNCIL OF CANADA
LE CONSEIL MÉDICAL DU CANADA

Table of Contents

List of Tables and Figures

Preface

This report summarizes the fundamental psychometric characteristics, test development and test administration activities of the Medical Council of Canada Qualifying Examination (MCCQE) Part I and candidate performance on the exam in 2017. Sections 1 to 5 describe the exam's purpose, format, content development, administration, scoring and score reporting. These sections also provide validity evidence in support of score interpretation, reliability and errors of measurement, and other psychometric characteristics. Section 6 summarizes candidate performances for the two administrations in 2017 and includes historical data for reference purposes. The report is intended to serve as technical documentation and reference materials for the Central Examination Committee (CEC), test committee members, Medical Council of Canada (MCC) staff, MCC stakeholders, and members of the public.

1. Overview of the MCCQE Part I

The MCCQE Part I is a one-day, computer-based exam that assesses the critical medical knowledge and Clinical Decision-Making ability of a candidate at a level expected of a medical student who is completing his or her medical degree in Canada. The examination is based on the MCC Objectives, which are organized under the CanMEDS roles, and covers the following specialty areas: Medicine, Obstetrics and Gynecology (OB/GYN), Pediatrics, Population Health and the Considerations of the Legal, Ethical and Organizational Aspects of the Practice of Medicine (PHELO), Psychiatry and Surgery.

The MCCQE Part I is composed of two components. The first component consists of 196 Multiple-Choice Questions (MCQs). The second component consists of 45 to 55 Clinical Decision Making (CDM) cases that include both short-menu and short-answer, write-in items.

The CEC is responsible for overseeing the MCCQE Part I including exam specifications, development of the exam, maintenance of its content and the approval of results.

2. Exam development

2.1 Exam specifications

2.1.1 The MCQ component

MCQs are single-correct answer-based items. MCQ exam specifications are limited to content covering the six specialty areas and control of difficulty levels within testlets. A testlet is a testing unit comprised of four MCQs of the same specialty area. Test committees generate testlets by level of difficulty ensuring that each testlet covers a variety of content for each specialty area. Testlets are used in the delivery of multi-stage adaptive testing (MSAT) of the MCQ component. Items are assigned a difficulty level based on their Rasch difficulty parameter established during calibration. The calibration process is described in Section 5.2. Four levels of difficulty are used: level 1 is comprised of very easy items; level 2 is comprised of easy items; level 3 is comprised of difficult items; and finally, level 4 is comprised of very difficult items. In Table 1, the mean difficulty

by specialty area and level of difficulty for 2017 is presented along with the minimum and maximum values by level of difficulty. A more comprehensive description of MSAT is covered in Section 5.4.

Table 1: Rasch difficulty parameter statistics by specialty area and levels of difficulty

Specialty Area(s)	Difficulty level	Mean	Min	Max
Medicine	1	-2.40	-5.01	-1.62
	2	-1.11	-1.62	-0.64
	3	-0.18	-0.63	0.30
	4	0.91	0.30	2.66
Obstetrics and Gynecology	1	-2.40	-4.87	-1.63
	2	-1.10	-1.62	-0.63
	3	-0.20	-0.63	0.30
	4	0.95	0.31	2.89
Pediatrics	1	-2.38	-4.51	-1.63
	2	-1.08	-1.59	-0.64
	3	-0.20	-0.64	0.32
	4	0.93	0.34	2.84
PHELO	1	-2.42	-4.97	-1.63
	2	-1.12	-1.63	-0.65
	3	-0.20	-0.63	0.30
	4	0.95	0.31	2.75
Psychiatry	1	-2.43	-4.93	-1.62
	2	-1.14	-1.62	-0.64
	3	-0.19	-0.63	0.32
	4	0.86	0.32	2.88
Surgery	1	-2.29	-4.95	-1.62
	2	-1.10	-1.62	-0.63
	3	-0.19	-0.63	0.31
	4	1.07	0.31	3.15

Percentages by specialty area are limited to PHELO. Based on weights that were decided before the implementation of computerized testing, Population Health was to constitute no more than 20 per cent of the PHELO content and the Legal/Ethical/Organizational component was to account for the remaining 80 per cent. When the MCC transitioned to computerized testing and MSAT

with its four-item testlets, the 20 per cent vs. 80 per cent ratio was translated to 25 per cent vs. 75 per cent. As such, Population Health was to contribute one item to a four-item testlet. The exam specifications are defined by the CEC, who ensures that exam content reflects the medical knowledge and Clinical Decision-Making ability at a level expected of a medical student who is completing his or her medical degree in Canada.

All content areas are weighted equally with the exception of PHELO, which is made of 25 per cent Population Health and 75 per cent Legal/Ethical/Organizational content.

2.1.2 The CDM component

The CDM component of the MCCQE Part I consists of six caselets that cover the specifications outlined in Table 2. A caselet is comprised of one case from each of the six specialty areas (i.e., six cases per caselet). Each case is comprised of one to four items that relate to the clinical case. Each form is comprised of six caselets for a total of 45 to 55 cases (including pilots). To control exposure within a multi-day exam period, eight CDM forms are assembled each year. The case distribution for the eight-form CDM component is presented in Table 3, which shows the overlap in content from one form to another. In this design, 108 unique cases are required for the eight forms in an administration.

Table 2: CDM eight-form caselet design

Priority 1: Complexity	
Cases and items of a more complex nature than MCQs	
Priority 2: Clinical tasks	per form
Data gathering	40%
Data interpretation	20%
Management	40%
Priority 3: Age group	per form
Pregnancy, perinatal, infant	10%
Pediatric (child and adolescent)	30%
Adult	45%
Geriatric	15%
Priority 4: Systems	
A variety of systems should be sampled	

Table 3: CDM eight-form caselet equating design

Caselet	FORM								Sum
	1	2	3	4	5	6	7	8	
1	X	X				X			**3**
2	X	X				X			**3**
3	X	X					X		**3**
4	X		X			X			**3**
5	X		X				X		**3**
6	X		X				X		**3**
7		X		X				X	**3**
8		X		X				X	**3**
9		X		X				X	**3**
10			X		X	X			**3**
11			X		X		X		**3**
12			X		X			X	**3**
13				X		X			**2**
14				X			X		**2**
15				X				X	**2**
16					X	X			**2**
17					X		X		**2**
18					X			X	**2**
Sum	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	

** Each "X" represents a set of six cases*

The Rasch model was adopted in the spring of 2015 and the unit of measurement became the item. Cases and caselets were not retained as measurement entities. However, caselet by specialty area design was retained to maintain control over content coverage.

2.2 Exam format

The exam consists of a MCQ component and a CDM component, each covering content in six specialty areas: Medicine, Obstetrics and Gynecology, Pediatrics, PHELO, Psychiatry and Surgery.

The MCQ component of the MCCQE Part I consists of seven sections, each composed of 28 items for a total of 196 items. The maximum time allotted for this component is three and a half hours. This component is designed as a multi-stage, semi-adaptive exam. This model allows for

an initial estimation of a candidate's ability following scoring of the first section (referred to as the routing section). Decisions are then made as to the level of difficulty of items in the next section. (A detailed description of the multi-stage model is covered in Section 5.4 of this report.) Each MCQ has a stem and five options, of which only one is the correct answer. There are no penalties for incorrect answers. The MCQ delivery model is designed in such a way that once candidates submit their answers to a particular section, they are not allowed to return to that section.

The CDM component consists of approximately 45-55 cases (including pilot items), with one to four items in each case, for a total of approximately 80 items. The maximum time allotted for this component is four hours. CDM items include both short-menu and short-answer, write-in formats. The CDM format is designed to assess problem-solving and Clinical Decision-Making skills. Candidates are presented with case descriptions followed by one or more test items that assess key issues in the resolution of the case. CDM items, as well as some MCQs, have pictorial material presented in the form of photographs, diagrams, radiographs, electrocardiograms and graphic or tabulated material. Candidates may be asked to elicit clinical information, order diagnostic procedures, make diagnoses or prescribe therapy. Their decisions should reflect the management of an actual patient.

Each candidate taking the CDM exam is assigned a test form at random. These forms are designed to include a set number of cases/items, evenly distributed across the six specialty areas. Within a test form, a candidate is also presented with approximately 10 pilot cases. Unlike the MCQ component, these pilot cases do not count toward a candidate's score. For cases containing items that perform well, they are banked as an active case for future use. If a repeating candidate is taking the exam twice within a given year, a different form is assigned to ensure they receive different cases

Typically, the MCQ portion of the exam is delivered in the morning and the CDM portion is delivered in the afternoon.

2.3 Item development

For the MCQ component, exam content is developed by each of the six specialty area-specific test committees that are comprised of family physicians and other specialists. Test committees include representation from both official language groups (English and French) as exam content is produced in both official languages. For the CDM component, exam content is developed by a

multi-disciplinary test committee with representation from each of the six specialty areas as well as from family physicians. The CEC Vice-Chair also sits on each MCCQE Part I test committee. The Vice-Chair contributes to the overall training of item authors, is a consistent member across committees, and supports the TDOs in identifying blueprint gap areas.

All new content from each MCQ and CDM test committee is reviewed and approved for piloting. For the MCQ component, new content is piloted before it is used as active content on the exam. The MCC analyzes candidates' response patterns after the exam. Pilot items that meet statistical criteria are included in the scoring. Pilot items that do not meet those statistical criteria are returned to their respective specialty area test committee for review and revision and are subsequently re-piloted. For the CDM component, new items or cases are piloted and scrutinized in a similar fashion; however, pilot items are used only for scoring on subsequent administrations when they meet performance requirements.

A total of 1145 MCQ items were piloted in the 2017 exam. Approximately, 70% of these items were developed using the traditional item writing method and 30% created using the Automated Item Generation (AIG) method.

It is standard practice to consider the purpose of the exam when preparing test items. Although the principle of developing MCQs and CDM items is similar, some differences exist. The following section outlines the item development cycle for MCQs and CDM items as well as the translation of items from English to French.

2.3.1 Multiple-Choice Questions

MCCQE Part I MCQ content is developed by six specialty area-specific test committees. Each committee is comprised of eight to 11 subject matter experts (SMEs) from across Canada who are experts in the fields of medical education and assessment. Test committees include representation from both official language groups and geographic representation from across Canada. Each test committee consists of a minimum of two family physicians. SMEs can be recommended by an MCC Test Development Officer (TDO), test committee member or by MCC's Selection Committee. All recommendations are approved by the Selection Committee at the MCC's Annual Meeting.

Each test committee meets for three days at least once per year at the MCC's head office in Ottawa. During these meetings, MCQs are written, classified, peer-reviewed and approved for

piloting. Content is developed by following professional standards outlined in Sections 3.1, 3.7, and 3.11 of the Standards for Educational and Psychological Testing (2014), as well as the guidelines outlined under 2.3 of the International Test Commission Guidelines on Test Use (2001). These standards and guidelines include quality assurance steps. First, subsequent to the test committees vetting and approving their items, the TDO and Examination Content Editors review the content for style, structure, and acceptable language appropriate for use in the exam. Second, the English version of the items are sent for translation. After translation, the MCC engages with the francophone universities to ensure language is inclusive of regional differences in Quebec. Lastly, the TDOs and Examination Content Editors complete an in-depth comparative read and validation of English and French items and then engage bilingual test committee members for an out-loud, comparative read of all items.

TDOs, in conjunction with the Chair of each test committee, guide test committee members to develop content where known content gaps in the exam specifications exist. Item development focuses on creating items with a range of difficulty levels, using most up-to-date medical terminology (for example, compliant with the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders [DSM-5]) and targeting items to meet exam specifications. Committee members are often asked to think about where the minimally competent candidate makes an error and use this as the focus in the development of items.

2.3.2 Automated item generation

In anticipation that the MCC would require larger numbers of test items, a three-year research project began in 2013 to explore the feasibility of implementing automated item generation (AIG) to develop MCQs. Test committees were introduced to the process of AIG in 2016.

AIG is a process by which cognitive models are used to generate items with computer technology using a three-step process (Gierl et al., 2013):

- **Step 1:** Medical experts identify and organize content for item generation. This content is used for the development of cognitive models
- **Step 2:** Medical experts create an item model that is used to specify where the cognitive model content must be placed in a template to generate items
- **Step 3:** Medical experts use a computer-based algorithm, the Item Generator (IGOR), to place content into the item model

IGOR is a JAVA-based software program developed to assemble the content specified in an item model, subject to the elements and constraints identified in the cognitive model. To improve user-friendliness, a web-based application, iButler (Medical Council of Canada, 2015), was developed in collaboration with two researchers from the University of Alberta. iButler allows test committee members to develop cognitive maps and generate items automatically on the fly. It is important to note that AIG, a tool to augment the development of items, will not replace traditional development of items.

AIG, using iButler, was launched operationally with MCCQE Part I test committees as of January 2016. The goal was to introduce and incorporate AIG as part of each test committee meeting with training on the development of new cognitive maps using the iButler software. All MCCQE Part I MCQ test committees were introduced to the concept of developing cognitive models using iButler. During each test committee meeting, a scheduled half-day session began with training on "what" AIG consists of followed by an interactive group exercise on how to create cognitive maps. Finally, a tutorial was provided on inputting the data/coding into the iButler software.

In 2017, the goal for each test committee meeting was to generate 80-100 items from a newly developed model and select the "best" 20 items for piloting on future MCCQE Part I forms. Generating this number of items enabled the committee sufficient sampling to choose a variety of AIG items from each model. Table 4 outlines the number of cognitive models developed and the number of items generated from these models.

Table 4: Number of items banked via AIG for each test committee in 2017

Test Committee	# of Models	# of Items
Medicine	3	40
Obstetrics/Gynecology	4	80
Pediatrics	3	60
PHELO	1	20
Psychiatry	5	100
Surgery	1	20
TOTAL	**20**	**320**

Overall, the feedback received from committees on this new AIG approach to developing MCQs was positive. AIG will be incorporated as part of regular ongoing activities to supplement traditionally developed items.

2.3.3 Clinical Decision-Making items

The CDM Test Committee is responsible for developing content for the CDM portion of the MCCQE Part I. This committee is comprised of SMEs from across specialty areas (Medicine, Obstetrics and Gynecology, PHELO, Psychiatry, Surgery and Family Medicine). The CDM Test Committee is structured with representation from the two official language groups and gender and geographic representation from across Canada. Similar to the content development of MCQs, content is developed by following professional standards mentioned in section 2.3.1 and rigorous quality assurance processes. Committee members meet twice per year and their mandate is to create, review and classify CDM content based on existing gaps.

The basis for the development of a CDM item is known as the key feature approach. This approach is based on the notion of case specificity, namely that clinical performance on one problem may not be a good predictor of performance on other problems. Consequently, assessments of clinical performance need to sample broadly as skills do not generalize across problems. To sample broadly in a fixed amount of time (four hours), assessment is best served by focusing exclusively on the unique challenges (i.e., key features) in the resolution of each problem, be they essential issues or specific difficulties. Test committee members are reminded to think about where the minimally competent candidate makes an error and use this as the focus for the development of key features.

The development of key feature-based cases for the CDM has been guided by psychometric considerations of content validity, test score reliability and sound principles of test development. Key feature cases provide flexibility on issues of item format (short-menu versus write-in), multiple responses to items and scoring criteria. Key feature problems have been found to be useful in assessments that require medical knowledge and the ability to apply that knowledge within clinical scenarios. These scenarios often require critical decisions to be made during the assessment and management of a given clinical scenario. These specific, critical decision points constitute the key features of the problem.

Once test committee members have created and approved key features, they continue with case development. At this point, the case and questions are developed in accordance with the scenario and selected MCC Objective. The CDM scoring keys reflect the main tasks that candidates must perform as identified in the key feature. All developed cases are approved by the test committee before being piloted. As an additional quality assurance step, content is vetted by the six MCQ specialty test committees and, if necessary, feedback is sent back to the CDM Test Committee suggesting content revision. Once a case has been piloted and has performed adequately, the case is banked as an available, "counting" case ready to be used on a future exam.

2.3.4 Translation of MCQs and CDM items

Exam items are initially created in English. The MCC then sends the items to professional translators with medical terminology translation expertise. Once the translation to French is complete, quality assurance steps are taken, and content revisions are made as required:

- The MCC's in-house editors perform a comparative read (comparing English items to French translations) of all items after translation is received and after each content review step (e.g. francophone member review, francophone university review, TDO review)
- A translation validation session is held where Francophone physicians from Francophone faculties of medicine participate in another round of comparative readings. Each French exam item is then reviewed by two to three Francophone physicians during these sessions.
- As a final step, a Francophone test committee member and an Examination Content Editor perform a final set of comparative reviews that include reading the content out loud and making final editorial content changes.

3. Exam Administration

3.1 Exam centres and exam delivery

The MCCQE Part I is offered twice per year in April/May and October/November during two- to

three-week testing windows at 26 sites, in both university computer labs and private testing centres across Canada.

The exam is delivered and monitored by MCC staff, through the QEI.net system developed by MCC Information Technology (IT) directorate. During the exam, site coordinators, who administer the exam at the faculties of medicine, are required to call in to MCC staff each morning to access security permissions to log into the exam. Each site coordinator has a personal identification code he or she must enter along with the candidate's code and personal identification number (PIN) for the exam to start. Site coordinators work directly with MCC staff to address technical permissions, security issues, technological issues and emergency situations.

The number of days a centre administers the MCCQE Part I depends on the maximum daily space capacity and the demand for that centre. The exam may be taken in either English or French at any centre; however, staff and technical support may be limited to a specific language. Support in both official languages occurs at the Ottawa and Montreal centres. A list of test centres is found in Appendix A.

3.2 Exam security

The MCC takes several measures to safeguard exam security. Test publishing processes are well established, test centre guidelines (exam delivery) are shared and reviewed with each site administrator prior to each testing window, and results processing is completed in the MCC's secure environment. This cycle of test delivery offers the MCC assurances of a consistent and fair exam administration for all candidates. The MCC collaborates with stakeholders on all facets of the exam process to ensure that only eligible candidates are allowed to write and that no one has an unfair advantage.

Every site administrator at each testing centre is trained to recognize potential test security breaches. Training occurs via site visits when new sites are opened or when there is a new site coordinator. The MCC follows up with verbal and written communication to update and reinforce security measures. In addition to test security at the test sites, MCC staff monitors online study forums for any candidate who may share exam content online before, during and after the administration.

3.3 Exam preparation

Online preparatory materials are available to help candidates prepare for the MCCQE Part I. These resources include demonstration videos, self-assessment tools, a list of resources by medical specialty area, and the MCC Objectives. All candidates have access to these materials through the MCC's website (*mcc.ca/examinations/mccqe-part-i/preparation-resources*). Additional support tools offered to candidates include the communication and cultural competence modules available through physiciansapply.ca. Preparatory tools will be enhanced in 2018 with the addition of new practice tests.

3.4 Quality assurance

After each exam administration, IT updates MCC's Post-CBT database with two basic SQL tables, namely one for each component of the exam. For each exam component there is a table that includes one row per item for each candidate. The tables contain the unique identifiers for candidates and items along with the candidate answers and scores for all counting and pilot items. An initial round of quality assurance of the tables is performed by the psychometrician for the MCCQE Part I, including a verification of completeness. Reasons for missing data are verified with the Evaluation Bureau. Once it is determined that the data meets the established quality assurance requirements, scoring and calibration are performed by Psychometrics and Assessment Services (PAS).

3.5 Release of results

Approximately five weeks following the last day of the exam session, the CEC meets to review performance on the exam, address administrative issues, rule on special candidate cases, and approve exam results. Table 5 outlines the specific statistical criteria the exam should meet to be approved and then have the results released. The MCC then grants candidates with access to their final result (such as pass/fail) and total score through their physiciansapply.ca account. Shortly thereafter, candidates have access to their Statement of Results (SOR), the official results document, and the Supplemental Feedback Report (SFR) that provides them with information on their strengths and weaknesses by specialty area and Clinical Decision Making.

Table 5 displays the statistical criterial for the approval of results for the 2017 exam. This table considers a total of 5,910 candidates, for the spring and fall sessions.

Table 5: Statistical criteria for the approval of results

Index		Best practice	Historical range[1]		Spring 2017	Fall 2017
			Spring	Fall		
Item performance	P-value	0.10 – 0.90[2]	0.02 – 0.99	0.02 – 0.99	0.05 – 0.99	0.05 – 0.99
	ITC	>0.30[2]	0.02 – 0.42	0.02 – 0.42	0.03 – 0.37	0.03 – 0.37
Decision accuracy		>0.90	0.90 – 0.96	0.84 – 0.91	0.93	0.89
Decision consistency		>0.90	0.91 – 0.94	0.84 – 0.85[3]	0.91	0.85
Pass rate (%)	CMG 1st	n/a	94.5 – 98.8	85.7 – 100	95.4	61.1
	Total	n/a	77.9 – 84.2	46.4 – 54.8	78.0	50.4

[1] *Based on 2012-2016 administrations.*

[2] *Items with flagged p-values or item-total correlations (ITCs) are reviewed by our Chief Medical Education Advisor, TDOs and TC members to rule out any content issue.*

[3] *Decision Consistency started being reported for fall administrations of 2015.*

4. Validity

"Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), 2014). Test validation requires gathering and integrating evidence from multiple sources to develop a validity argument that supports intended uses and interpretations of scores and to rule out threats to validity (Messick, 1989, 1994).

The validation of the MCCQE Part I is an ongoing process of gathering evidence in support of the interpretation of exam scores as one of the indicators of a candidate's basic medical knowledge and skills in the principal specialty areas of medicine. Validity considerations have been incorporated into exam design, exam specifications, item development, exam assembly, psychometric quality, exam administration and results reporting.

4.1 Evidence based on exam content

During the course of exam content development, great care is taken to ensure the exam is relevant to medical graduates entering postgraduate training in Canada. As indicated in Section 2,

MCCQE Part I items are developed based on exam specifications defined by the CEC members who ensure that exam assesses the critical medical knowledge and Clinical Decision-Making ability of a candidate at a level expected of a medical student who is completing his or her medical degree in Canada.

Various test committees are involved in developing test items. Regular content development workshops for each specialty area are conducted to train test committee members to develop items that reflect the knowledge and skills emphasized in the exam specifications for each content area and meet professional test development guidelines. The MCC's guidelines for item development have been documented and are available online. Guidelines have been developed for both MCQs and CDMs. The MCQ guidelines can be found here: *mcc.ca/wp-content/uploads/Multiple-Choice-Question-guidelines.pdf* and the CDM guidelines can be found here: *mcc.ca/wp-content/uploads/CDM-Guidelines.pdf*. The items are reviewed, edited and finalized by test committee members, TDOs, editors, and translators.

4.2 Evidence based on the exam's internal structure

As each candidate receives a comparable but different set of items, factor analysis is difficult to conduct on the MCCQE Part I as it requires more exam data than is available; however, the internal structure of the MCCQE Part I can be revealed, to some degree, through the evaluation of the correlations among specialty area subscores. This can help one understand how closely the exam conforms to the construct of interest. These correlations were examined using the data from 4,348 examinees who took the spring 2017 MCCQE Part I and had a final result of pass or fail.

Table 6 displays a correlation matrix of subscores in the six specialty areas covered by the exam. These correlations were corrected for attenuation, indicating what the correlation would be if we could measure each specialty area with perfect reliability.

One can observe that content domains correlate from moderately high (such as 0.73 between Medicine and PHELO) to high (such as 0.89 between Surgery and Pediatrics). This suggests that performance in the different content domains of the MCCQE Part I reflect an essentially single dominant underlying construct (for example, basic medical knowledge and clinical skills that the MCCQE Part I is designed to measure).

Table 6: Correlations (corrected for attenuation) among specialty areas (N = 4,348)

	Medicine	Obstetrics and Gynecology	Pediatrics	Surgery	Psychiatry
Obstetrics & Gynecology	0.82*				
Pediatrics	0.87*	0.87*			
Surgery	0.87*	0.87*	0.89*		
Psychiatry	0.78*	0.79*	0.82*	0.78*	
PHELO	0.73*	0.77*	0.77*	0.76*	0.87*

** Significant at p<0.001*

4.3 Minimizing construct-irrelevant factors

Another way to enhance validity is through the minimization of construct-irrelevant variance (for example, error variance caused by factors unrelated to the construct measured by the exam). During development, items are reviewed by SMEs and TDOs to ensure they meet the exam specifications. As well, SMEs and TDOs review items for appropriateness of language and potential, unintended bias against certain language or culture groups. In addition, empirical evidence from the item and distractor analysis is used to further investigate potential sources of construct irrelevance. This topic is further developed in Section 5. Test completion rates, candidate item response times and overall test times are also analyzed to ensure that time allotted to complete the exam is adequate and that speededness is not a factor affecting candidate performance. The MCC ensures that testing conditions across all test centres are standardized so that candidates have equal opportunities to demonstrate their abilities. Finally, detailed test information and links to resources are provided on the MCC's website to help candidates prepare for the exam and alleviate test anxiety.

5. Psychometric analyses

In this section, we describe the psychometric analyses completed following each exam administration. We conduct item analyses, followed by item calibration, estimation of candidates' ability, scoring, standard setting and scaling, and finally, score reporting.

5.1 Item analysis: Classical test theory and item response theory

Following each administration of the MCCQE Part I, the PAS team conducts item analyses to verify the soundness of each item from a statistical perspective prior to engaging in final scoring of the exam. Item analysis, using both classical test theory and item response theory, results in items being flagged for various reasons outlined below. The inclusion or exclusion of items flagged during item analysis in final scoring is predicated on a careful content review by experts. While content experts are encouraged to use the statistical information in the review process, the final decision rests on whether the content is defensible given the intent of the item and/or case.

Classical test theory flags

Immediately following an administration, an initial item analysis (IIA) is conducted using responses from all first-time test takers. An IIA involves a classical item analysis to review item difficulty, discrimination, and candidate raw-score performance. Specifically, p-values are computed as a measure of an item's difficulty and an item-total correlation is computed to reflect item discrimination. A point-biserial correlation is computed for dichotomously scored items such as MCQs (items scored 0 or 1) and a polyserial correlation is computed for polytomously (more than two score categories) scored items such as CDM write-ins (items with more than two score categories, for example, 0, 0.33, 0.67 and 1). In addition, PAS examines the proportion of candidates who select each option as an indicator of how well each distractor (the incorrect responses) is functioning. The investigation of how well each distractor is performing is supported by computing the correlation between each distractor and the total score. If distractors are performing as intended, these correlations will be negative (for example, candidates with lower overall MCCQE Part I scores are selecting the distractors more frequently than higher-ability candidates).

Items flagged by PAS are reviewed by both psychometricians and content experts. An item is flagged if it meets one or more of the following rules:

- Very high difficulty: p-value<0.10
- Very low difficulty: p-value>0.95
- Item mean square outfit: >2
- High percentage of omits >5 per cent
- Low correlation value for the correct answer: <0.05

- Distractor correlation is positive and the magnitude of the statistic is stronger than for the correct answer

Flagged items are included in final IRT calibrations only after psychometricians and content experts have reviewed the items and confirmed that the content is acceptable and the key is correct. Items flagged during IIA and determined to be flawed after review will be removed from further analyses with the review committee's approval.

5.2 IRT item calibration

Previous research studies (De Champlain, Boulais, & Dallas, 2012; Morin, Boulais, & De Champlain, 2014) have established that simpler models, such as the Rasch model, yield results that are consistent with those from more elaborate models such as the two-parameter logistic model. Starting with the spring 2015 administration, the Rasch model and one of its extensions, the partial credit model (Masters, 1982), were applied, using Winsteps (Linacre, 2015), to the MCCQE Part I for item calibration and scoring. This transition has allowed the implementation of a unified IRT model for the estimation of all MCQ and CDM dichotomous and polytomous items as well as establishing candidate abilities by considering all items together (MCQs and CDMs).

With the Rasch model, the probability of a correct response on a dichotomous item is modeled as a logistic function of the difference between the ability of a person and the item difficulty parameter. If X = 1 denotes a correct response and X = 0 denotes an incorrect response, for the Rasch model, the probability of a correct response takes on the following form:

$$P_i\{X_{ni}\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}},$$

where β_n is the ability of person *n* and δ_i is the difficulty of item i.

For polytomous items, the polytomous Rasch model (partial-credit model) is a generalization of the dichotomous model. It is a general measurement model that provides a theoretical foundation for the use of sequential integer scores (categorical scores).

For the spring 2017 MCCQE Part I, items with banked Rasch difficulty parameter estimates were used as fixed values to calibrate new, freely estimated items, following the steps outlined below.

For calibration purposes, a reference group comprised of Canadian and international medical graduates first-time test takers is used. Therefore, all repeat test takers are excluded from the first four steps of calibration.

- **Step 1:** The goal of the first step is to estimate the MCQ item parameters for the pilot items and to identify potential 'poor performing' items. The first run consists of calibrating the MCQs, based on the responses of the first-time test takers, with the item difficulty parameters of the counting items set as anchor values and the item difficulty of pilot items freely estimated. Through this step, items that do not satisfy the statistical criteria outlined in Section 5.1 are flagged and are to be reviewed by subject matter experts. The decision to be made is to retain or remove from scoring.
- **Step 2:** After the TDO makes arrangements with the subject matter experts to review all flagged items (in Step 1) and to decide which items will be removed from scoring and calibration, the pilot MCQs are then recalibrated. A final set of calibrated items are then ready to be used in step 3.
- **Step 3:** All CDM dichotomous and polytomous items are calibrated using calibrated MCQs as anchors (or fixed values). A content review of flagged CDM items is done following this step.

Since the adoption of the Rasch IRT model for the calibration and scoring in the spring 2015 MCCQE Part I, additional statistical criteria have been introduced for the CDM component to identify potentially flawed items.

Currently, the CDM component has dichotomous as well as polytomous items. For polytomous items, an extension of the Rasch model, the partial credit model, is used to establish the difficulty level that takes into account step parameters or step thresholds. These thresholds are model-based and are assumed to increase in value as the score categories increase. It is expected that candidates' average abilities advance across categories for CDM items; a score of 0.67 on an item requires higher overall ability than a score of 0.33. When this expectation is not met, these items are referred to as having disordered step parameters (for instance, weaker candidates overall on the exam obtain higher scores on the item than abler candidates). These items are flagged as potentially flawed and subject to content review.

Furthermore, polytomous items with near zero option endorsement (for example, too few candidates who obtain a particular score) are also flagged for content review.

Finally, CDM write-in items that display low inter-rater marking agreement are also flagged. It is expected that scores assigned by two markers would yield highly comparable results. CDM write-in items that display less than 90 per cent agreement between markers are flagged for review. Additionally, items that have weighted kappa coefficients less than 0.61 are also flagged for review. The kappa coefficient reflects the agreement between markers above and beyond chance agreement (Cohen, 1979).

Following the IIA in spring 2017, 132 MCQs were flagged and after consultation with a content expert, they were not included in the final scoring. For the CDM items, 56 were flagged and following consultation with a content expert, 11 were not included in final scoring.

- **Step 4:** CDMs are recalibrated following the exclusion of items identified in Step 3 to obtain final difficulty parameter estimates for all MCQs and CDMs.

5.3 Estimating candidate ability

Winsteps (Linacre, 2015) allows the user to calibrate items and estimate candidate abilities at the same time, using an iterative process and two estimation procedures (the PROX procedure, which is the Normal Approximation Algorithm devised by Cohen (1979), and a Joint Maximum Likelihood Estimation (JMLE) procedure). Estimates of item difficulty and candidate ability are obtained through an iterative process. Initially, all unanchored parameter estimates (measures) are set to zero. Next, the PROX method is employed to obtain rough estimates of item difficulties. Each iteration through the data improves the PROX estimates until they reach a pre-set statistical criterion. Those PROX estimates are the initial estimates for JMLE, which fine-tunes them again by iterating through the data to obtain the final JMLE estimates. This iterative process ceases when the convergence criteria are met. In Winsteps, two convergence criteria can be set to establish stopping rules for the iterative process (Linacre, 2016). For high precision, the logit (log-odds units) change criterion was set at 0.000001 and the residual score criterion was set at 0.0001.

- **Step 5** yields person abilities using all MCQ and CDM items post-step 4.

Given that the same MCQs and CDMs are used in the fall and the spring, ability estimates in the fall administration are obtained by using the same item parameter estimates as established in the last calibration step from the spring administration.

5.4 Multi-stage adaptive test delivery

After several years of research, computer-based testing (CBT) was first introduced in the fall 2000 MCCQE Part I administration. Maguire (1999) established that the two-parameter logistic model was the best fit given item responses to the MCQ component of the MCCQE Part I. It was further established (Maguire, 2001) that there was a significant, high correlation between the total number correct and ability estimates as calculated using the two-parameter logistic IRT model. Along with the decision to use the total number of items (per medical specialty area) as a stopping rule (for instance, end of the exam), traditional adaptive testing was replaced with a version of multi-stage adaptive testing whereby a routing section is used to route candidates to an appropriate level of difficulty given their responses to one set of four items (testlet) per specialty area.

Figure 1 outlines the logic implemented following the administration and scoring of the routing section. The first section of a seven-section MCQ component is therefore composed of routing testlets. In each specialty area, a routing testlet is comprised of four items of varying levels of difficulty (for example, one very easy item, one easy item, one difficult item and one very difficult item). After the answers to all items of the routing section are submitted, testlets are scored on the fly and decisions are made for each of the six specialty areas as to what level of difficulty the items will be in the second section. Starting with the second section, each specialty area testlet contains four items of the same difficulty level. A candidate who scores zero or one out of four items from the routing section will be presented with a testlet containing four level-one items in the second section (for example, four very easy items). A candidate who scores two out of four in a testlet of the routing section will be presented with a level-two testlet in the second section (for example, four level-two items). A candidate who scores three out of four in a routing testlet will be presented with level-three items in section two. Finally, a candidate who scores four out of four in a routing testlet will be presented with four level-four items in section two.

Sections two through six decision rules follow the same logic (see Figure 2). For example, a candidate who scores zero or one out of four items in a testlet of section two will be presented, in section three, with four items from one level downwards. A candidate who scored zero or one in a testlet from level four in section two will be presented with four items of level three in section three. If a candidate scores zero or one in a testlet from level one in section two, this same candidate will be presented with four items of the same level in section three, namely four items of level one.

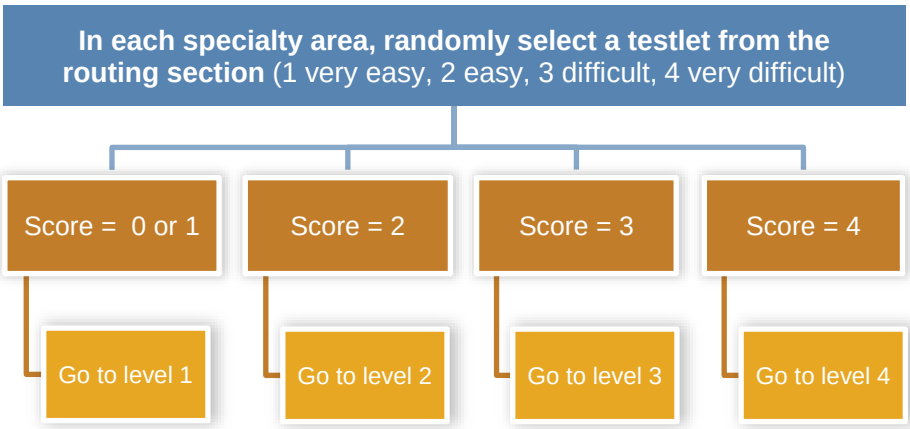


Figure 1: Multi-stage adaptive testing – routing section

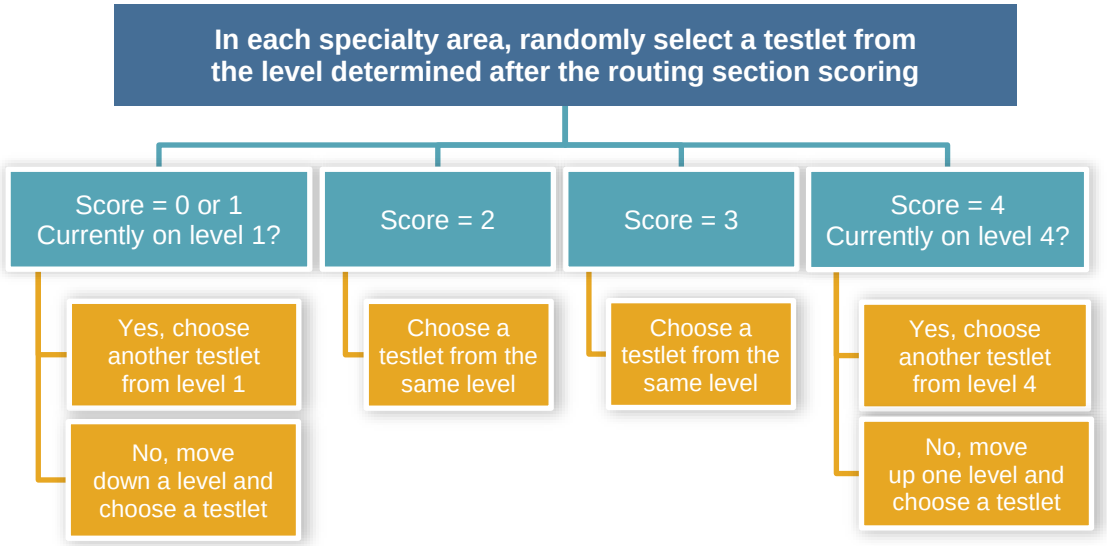


Figure 2: Multi-stage adaptive testing – sections 2 to 6 decisions

5.5 Scoring

A candidate's ability and total score on the MCCQE Part I is derived from combined performance on the MCQ and CDM components. The MCC uses the Rasch model (Rasch, 1960) to score candidates' exam responses. While raw score data (scores of the 1/0 type) are necessary, they are insufficient to establish a candidate's ability level. Simply adding up item scores does not

accurately reflect a candidate's ability since this does not take into account the difficulty level of the items that were encountered in any given MCCQE Part I form.

MCQ and CDM short-menu items are machine-scored as they involve numbered responses that are then compared to pre-defined scoring keys. CDM write-in items are marked by physician markers. Since the fall 2014 MCCQE Part I, physician markers have used the MCC-developed software application "Aggregator" to facilitate the marking of CDM constructed response items. Using the Aggregator, physician markers are presented with CDM cases, items, key features and scoring keys. Prior to being presented the answers, the Aggregator combines identical answers given by candidates for a given item. All unique answers that do not aggregate are also presented. Physician markers are then asked to indicate whether an answer is deemed correct or incorrect given pre-determined scoring keys (such as correct answers). Each item is marked independently by two physician markers and when discrepancies are detected, the issue is resolved by a third marker. The Aggregator also allows physician markers to indicate whether candidates have exceeded the number of answers allowed for an item. Markers do not assign scores to items; they are simply asked to indicate whether answers are correct or incorrect and scoring is performed following this validation step. Once all answers have been categorized as either correct or incorrect, scoring is done automatically, taking into account all other constraints such as exceeding the maximum number of answers allowed.

All MCQs are dichotomously scored as they all have one correct answer. A large proportion of CDM items is also dichotomously scored (70 per cent of counting items in 2017). For polytomous CDM items that involve more than one correct answer, the first step is to assign proportional scores. The second step is to assign categorical scores to each of the possible combination of proportional scores as these are the type of data that can be analyzed by the partial-credit model. For example, a candidate selecting two out of three correct answers would receive two-thirds of a mark (such as 0.67), that is then also assigned a categorical score of three out of four.

The Rasch model requires that each item's difficulty level be determined to assess a candidate's ability. The Rasch model (and an extension of this model, the partial-credit model that can handle CDM items that have more than one correct answer) allows us to establish a candidate's ability by considering the level of difficulty of all items. The Rasch model also allows us to establish a scale that is expressed in such a way that candidate attributes, such as ability, and item attributes such as item difficulty are on the same unit of measurement. In its initial phase, a scale is defined in measurement units called logits (log-odds units) and allows for candidates' abilities to be

expressed on the same scale as the item difficulties. Values typically range between -3.00 and +3.00 although values beyond the latter can occur. A candidate who obtains a score of -3.00 would demonstrate very little knowledge in regards to the specialty areas being assessed whereas a candidate who obtains a score of +3.00 would demonstrate strong knowledge.

5.6 Standard setting and scaling

The MCC conducts a standard-setting exercise every three to five years to ensure the standard and the pass score remain appropriate. Standard setting is a process used to define an acceptable level of performance and to establish a pass score.

In the fall of 2014, the MCC completed a rigorous standard-setting exercise based on expert judgments from a panel of 17 physicians representing faculties of medicine from across the country, different specialties and years of experience supervising students and residents.1 The Bookmark Method, a successfully employed and defended method used by large-scale exam programs, was used for the to help panelists to suggest a new pass score for the exam. The panelists selected for a standard-setting exercise represent a microcosm of all MCCQE Part I examination stakeholders and were representative with respect to a number of key variables, including the region of Canada, ethnicity, medical specialty and years of experience. The recommended cut score was subsequently brought forward to the CEC for consideration and approval. The CEC, whose members are appointed annually by the MCC's Council, is responsible for the quality of MCC examinations and awards final results, such as pass or fail, to candidates.

In the spring 2015 MCCQE Part I, a new standard was applied to reflect this minimally-acceptable level of performance. The value representing this standard was established at -0.22 on the Rasch scale. Though the Rasch scale defined above has properties that are well suited for mathematical calculations, it is not very user-friendly for the candidate population. A linear transformation of the Rasch ability estimate is necessary to establish a scale of reported scores that is more meaningful to candidates. The scale chosen has a mean of 500 and a standard deviation of 100. On that scale, the pass score is equivalent to 427 for the MCCQE Part I.

[1] *mcc.ca/wp-content/uploads/MCCQE-Part-I-Standard-Setting-Report-2015.pdf*

To establish an individual candidate's scale score, a linear transformation is performed. The following generic formula is applied:

$$X_i' = a + bX_i$$

Where $X_i' =$ scaled score;

$b =$ the multiplicative component of the linear transformation often referred to as the slope;

$a =$ the additive component often referred to as the intercept;

And $X_i =$ a candidate's Rasch ability score

In the spring of 2015, when the scale was first established, the slope and intercept were established to be 215.7309 and 475.0214 respectively. These two constants were applied to transform each candidate's Rasch ability score into a scale score.

A candidate's final result such as pass or fail is determined by his or her total score and where it falls in relation to the exam pass score; a total score equal to or greater than the pass score is a pass and a total score less than the pass score is a fail. The candidate's performance is judged in relation to the exam pass score and not judged on how well other individuals perform.

5.7 Score reporting

Approximately seven weeks after the last day of the exam session, the MCC issues a Statement of Results (SoR) and a Supplemental Feedback Report (SFR) to each candidate through their physiciansapply.ca account. Samples of the SoR and SFR can be found in Appendix B. The SoR

includes the candidate's final result and total score as well as the score required to pass the exam. Additional information about specialty area, CDM subscores, and comparative information is provided in the SFR, offering the candidate information on areas of strengths and weaknesses. Since subscores have fewer items, there is less measurement precision. Subscores are provided to individual candidates for feedback only and are not meant to be used by organizations for selection.

After the administration of an exam, a candidate whose performance has potentially been affected by procedural irregularities that occurred during that exam, is reported to the CEC for a special ruling. A candidate may receive a No Standing as the CEC cannot, in these cases,

establish a valid pass or fail decision. In other special cases, such as candidates having been observed violating the exam's regulations (for example, having been observed using a smartphone during the exam), the CEC may award a Denied Standing.

6. Exam results

Candidate performance for the two administrations in 2017 is summarized in this section. When applicable, historical data from previous years are included for reference.

6.1 Candidate cohorts

In 2017, the MCCQE Part I was administered in a three-week window (April 24 to May 12) in the spring and in a two-week window (October 30 to November 10) in the fall. A total of 5,910 candidates challenged the exam across the 21 testing sites. Of the total number of candidates who took the examination in 2017, 11 candidates received a No Standing. Five of the 11 No Standing candidates are from the spring session and were approved by the CEC on June 6, 2017. In the fall session, six candidates received a No Standing, as approved by CEC on November 30, 2017 and based on candidates who responded by December 31, 2017 to accept the No Standing. Table 7 summarizes the distribution of candidates across groups defined by their country of graduation and number of times they have written the MCCQE Part I.

Table 7: Group composition – 2017

	Spring 2017		Fall 2017		Total	
Group	N	%	N	%	N	%[1]
CMG first-time test takers	2784	64.0	18	1.2	2802	47.4
CMG repeat test takers	43	1.0	113	7.3	156	2.6
IMG first-time test takers	864	19.8	815	52.3	1679	28.4
IMG repeat test takers	662	15.2	611	39.2	1273	21.5
TOTAL	**4353**		**1557**		**5910**	

[1] Percentages do not total 100 due to rounding.

6.2 Overall exam results

Table 8 summarizes pass rates for the 2017 spring and fall cohorts as well as for the whole year, along with basic descriptive statistics. The scores are presented on the reporting scale, which ranges from 50 to 950; the pass score is 427. Table 8 does not include the 11 candidates who received a No Standing.

Table 8: Exam results – spring and fall 2017

		Exam Results		
		Spring 2017	Fall 2017	Total
CMG First-time Test Takers	N	2784	18	2802
	M	546	455	545
	SD	69	71	69
	Min	305	354	305
	Max	765	570	765
	Pass Rate (%)	95	61	95
CMG Repeat Test Takers	N	43	113	156
	M	428	451	445
	SD	59	48	52
	Min	290	281	281
	Max	605	564	605
	Pass Rate (%)	49	69	63
IMG First-time Test Takers	N	863	814	1677
	M	453	451	452
	SD	95	96	95
	Min	50	50	50
	Max	794	767	794
	Pass Rate (%)	62	63	62
IMG Repeat Test Takers	N	658	606	1264
	M	385	389	387
	SD	68	67	68
	Min	50	173	50
	Max	561	578	578
	Pass Rate (%)	27	31	29
All Candidates	N	4348	1551	5899
	M	502	427	482
	SD	97	88	100
	Min	50	50	50
	Max	794	767	794
	Pass Rate (%)	78	51	71

Figure 3 displays the total score distribution on the reported score scale for all candidates in the spring, fall and total. Overall, the total score performance for the fall cohort was lower than for the spring cohort.

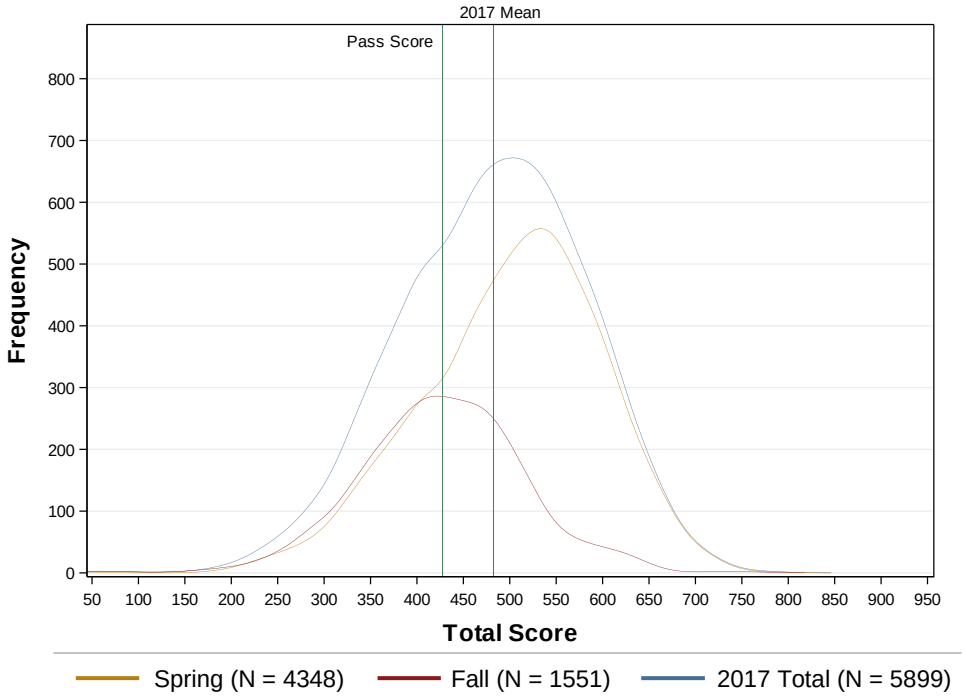


Figure 3: Total exam score distributions – spring and fall 2017

6.3 Reliability of exam scores and classification decisions

Test reliability refers to the extent to which the sample of items that comprises any exam accurately measures the intended construct. Reliability of the MCCQE Part I can be assessed by examining the standard error (SE) along the reported score scale. The SE indicates the precision with which the scores are reported at a given point on the scale and is inversely related to the amount of information provided by a test at that point. The SE values should be as small as possible so that the measurement of the candidate's ability contains as little error as possible. In the framework of IRT, the SE serves the same purpose as the standard error of measurement (SEM) in classical measurement theory (Hambleton, Swaminathan & Rogers, 1991), except that the SE varies with ability level in IRT whereas the classical SEM does not.

Figures 4 and 5 display scatter plots of SE values along the reported score scale for the spring and fall 2017 administrations, respectively. For each cohort, the plot shows that scores are less accurate toward the lower and higher ends of the score scale, but more accurate in the middle range of the scale where the majority of the scores fall. The SE is the lowest near the pass score, which indicates the highest precision of ability estimates, thus supporting more accurate and consistent pass/fail decisions.

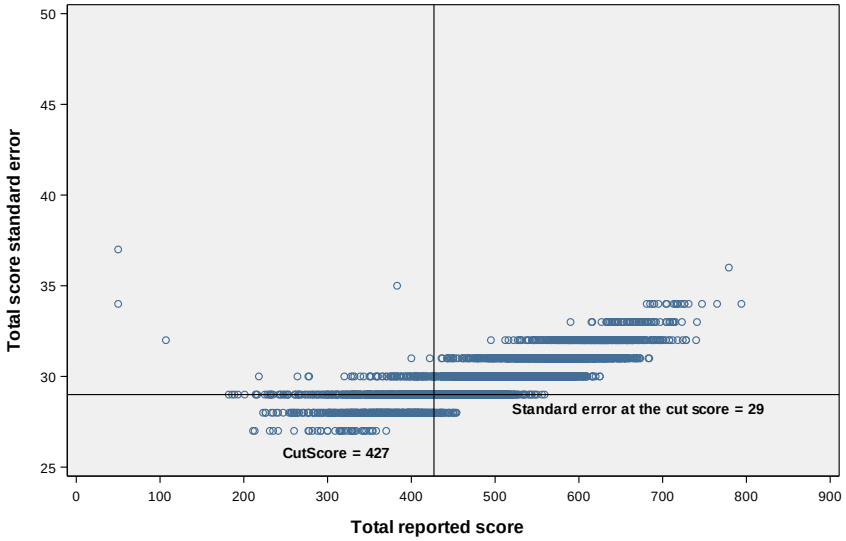


Figure 4: Total exam standard errors of ability – spring 2017

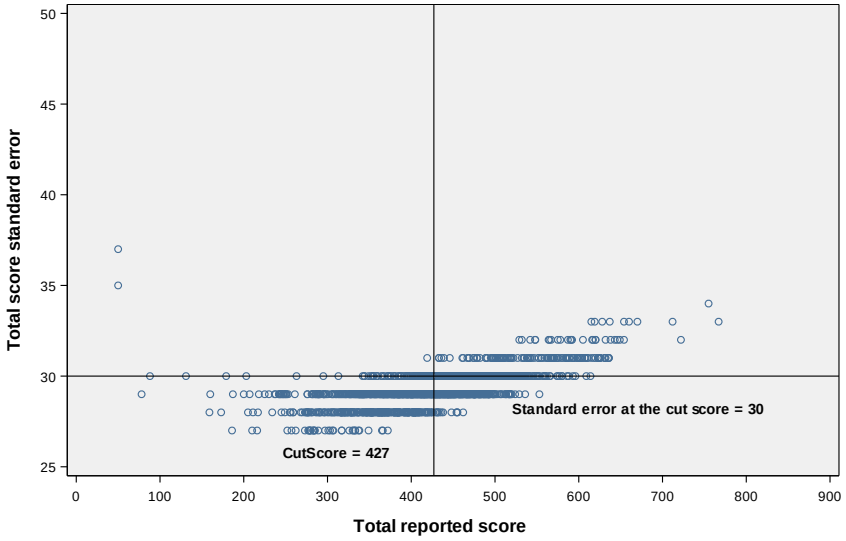


Figure 5: Total exam standard errors of ability – fall 2017

6.4 Pass/fail decision accuracy and consistency

In the context of this high-stakes exam, the accuracy of pass/fail decisions is of the utmost importance. Reliability of the MCCQE Part I can also be assessed by examining the consistency and accuracy of pass/fail decisions based on exam scores. Decision consistency and decision accuracy can be estimated using the Livingston and Lewis (1995) procedure that is used by many high-stakes testing programs. Decision consistency is an estimate of the agreement between pass/fail final decisions on potential parallel forms of the exam. Decision accuracy is the estimate of the agreement between the pass/fail decisions based on observed exam scores and those that would be based on their true score (for example, if the candidate could be tested on an infinite number of MCCQE Part I items). As indicated in Table 9, both the decision consistency estimate and the decision accuracy estimate for each of the two administrations of 2017 indicate reliable and valid pass/fail decisions based on MCCQE Part I scores. Table 9 is based on data from 4348 candidates in the spring session and 1551 candidates in the fall session.

Table 9: Reliability estimates, standard errors of measurement, decision consistency and decision accuracy indices for each administration of 2017

	Spring	Fall
Reliability estimate	0.90	0.88
SEM (score scale)	28.8	29.5
Decision consistency	0.91	0.85
False positive	0.05	0.08
False negative	0.05	0.08
Decision accuracy	0.93	0.89
False positive	0.03	0.05
False negative	0.04	0.06

6.5 Domain subscores profiles

The purpose of the domain subscore profile is to provide diagnostic information to candidates by highlighting their relative strengths and weaknesses. The SFR is designed to provide subscore information at the candidate level. In this report, we present domain subscore information for all candidates for the spring and fall 2017 administrations. The range of domain subscores is presented graphically in Figures 6 and 7. The graphs show the domain subscore for each of the eight domains. The boxes for each domain indicate the range of scores for 50 per cent of the

candidates' domain subscores. The vertical line represents the median or 50th percentile subscore. The remaining 50 per cent of domain subscores are shown to the right or the left of the box as a line (25 per cent to the right and 25 per cent to the left).

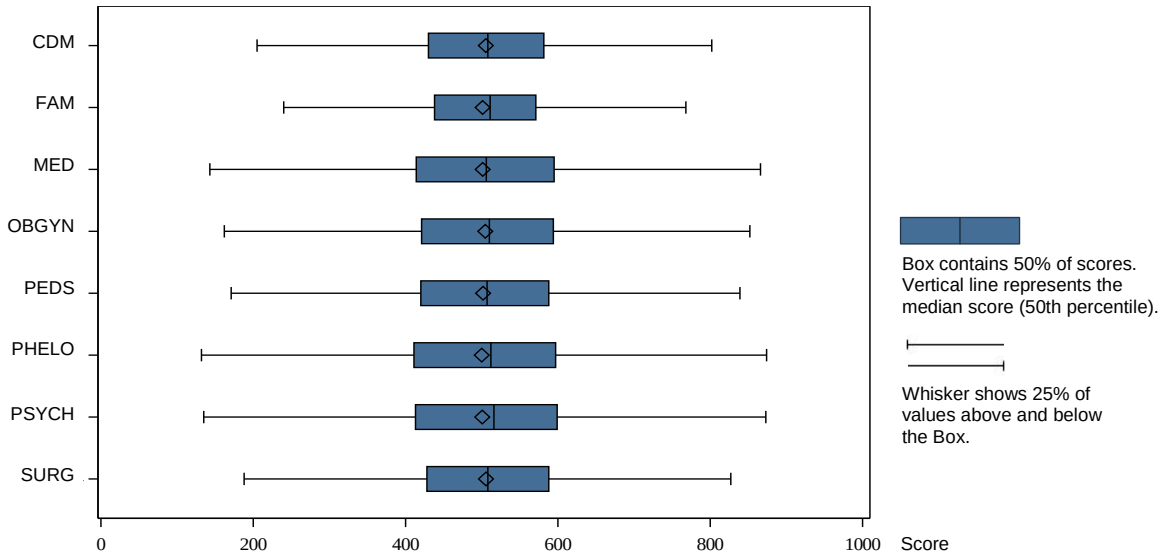


Figure 6: Domain subscore profile for the spring MCCQE Part I candidates

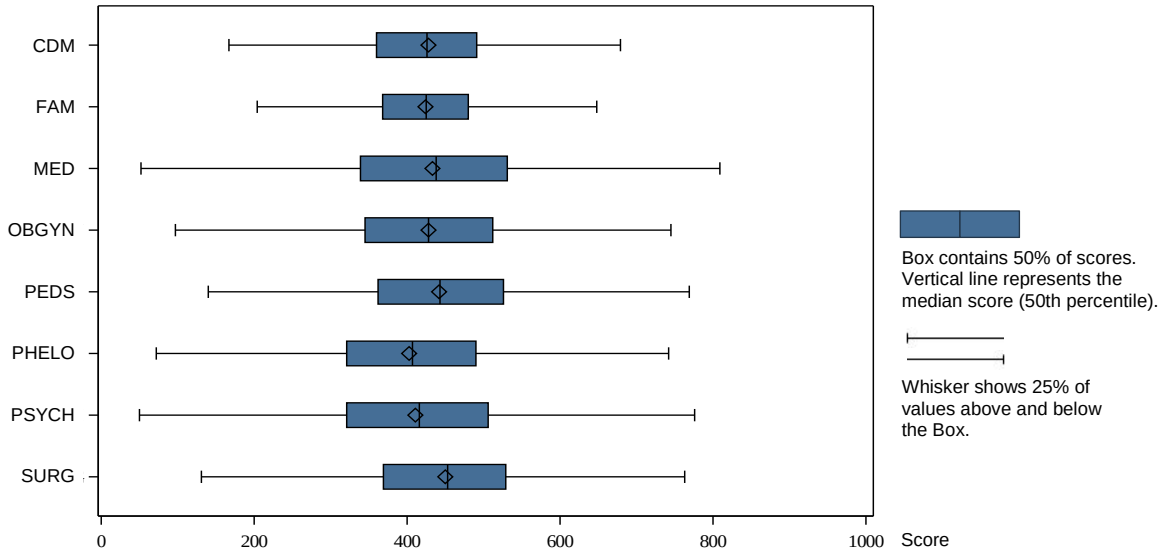


Figure 7: Domain subscore profile for the fall MCCQE Part I candidates

6.6 Historical pass rates

Historical pass rates are presented in this section. Table 10 shows the pass rates for 2015 to 2017 by group.

Table 10: Spring 2015 to fall 2017 pass rates

	2015		2016		2017	
	N	Pass rate	N	Pass rate	N	Pass rate
CMG first-time test takers	2791	95	2831	97	2802	95
CMG repeat takers	168	64	171	69	156	63
IMG first-time test takers	1638	60	1704	58	1677	62
IMG repeat takers	1023	20	1210	29	1264	29
TOTAL	**5260**	**70**	**5916**	**71**	**5899**	**71**

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1978). *A rating formulation for ordered response categories*. Psychometrika, 43, 561-73. *dx.doi.org/10.1007/BF02293814*.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington DC: American Council on Education.

Cizek, G. J. (2001). (Ed.). *Setting Performance Standards: Concepts, Methods and Perspectives*. New Jersey: Lawrence Erlbaum Associates Inc.

Cohen, Leslie. (1979). Approximate Expressions for Parameter Estimates in the Rasch Model. *The British Journal of Mathematical and Statistical Psychology*, 32, 113-120. *onlinelibrary.wiley.com/doi/10.1111/j.2044-8317.1979.tb00756.x/abstract*.

De Champlain, A., Boulais, A.-P., & Dallas, A. (2012). *Calibrating the Medical Council of Canada's Qualifying Part I Exam Using an Integrated Item Response Theory Framework: A Comparison of Models and Calibration Designs*. Ottawa, Canada: Medical Council of Canada. *dx.doi.org/10.3352/jeehp.2016.13.6*.

Gierl M, Lai H, Turner, S. (2012) *Using automatic item generation to create multiple-choice test items*. Medical Education, 46, 757-765. *onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2012.04289.x/abstract*.

Gierl, M.J., & Haladyna, T. (2013). *Automatic item generation: Theory and practice*. New York: Routledge.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson and J. S. Helmick (Eds.). *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass.

International Test Commission (2001). International Guidelines for Test Use, International Journal of Testing, 1(2), 93-114.

Linacre, J. M. (2015). *Winsteps® Rasch Measurement Computer Program*. Beaverton, Oregon: Winsteps.com.

Linacre, J.M. (2015). *Winsteps® (Version 3.91.0) [Computer Software]*. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2015. Available from *winsteps.com*.

Linacre, J. M. (2016). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.

Linacre J.M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, 16 (2) p.878. Retrieved from *rasch.org/rmt/rmt162f.htm*.

Livingston S.A. & Lewis C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179–197. *jstor.org/stable/1435147*.

Maguire, T. O., (2001) *Item and Testlet Use for the Multiple-Choice Portion of the May 2001 Qualifying Exam*. Unpublished paper.

Maguire, T.O. (1999). *Adaptive Testing and Part I of the Medical Council of Canada's Qualifying Exam*. Research and Information Report 1999-02.

Maguire, T.O. (2000). *Procedures for Calculating Equating Expressions and Standard Errors for the CRS Practice Exam*. Research and Information Report 2000-02.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174. *dx.doi.org/ 10.1007/BF02296272*.

Medical Council of Canada (2015). *iButler® (Version 1.3) [Computer Software].* Ottawa*,* Ontario.

Messick, S. (1989). Validity. In Educational Measurement (3rd ed., p. 610). Macmillan USA.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Morin, M., Boulais, A-P., & De Champlain, A. (2014) Scoring the Medical Council of Canada's Qualifying Exam Part I: A comparison of multiple IRT models using different calibration methods. Unpublished paper.

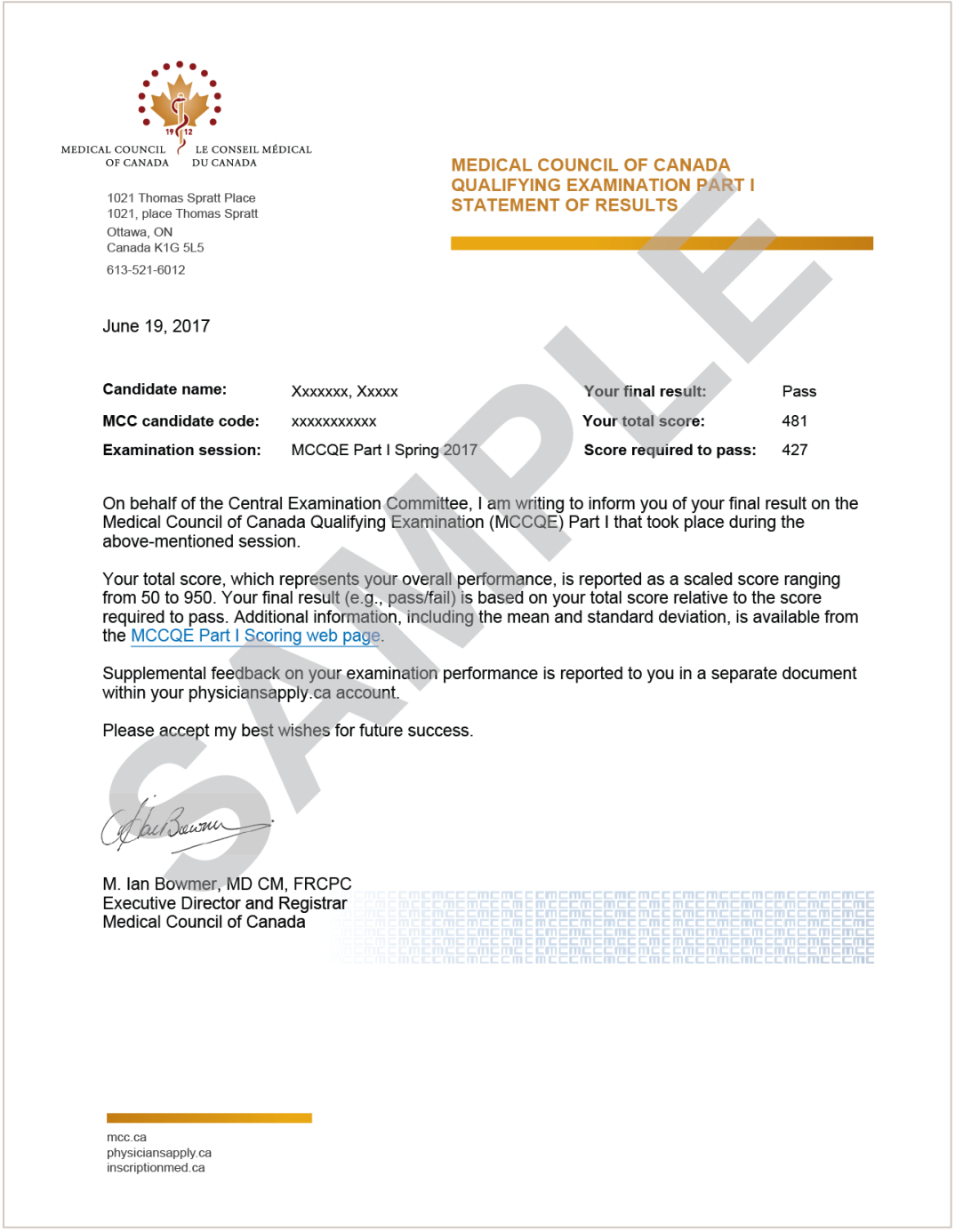
Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Smith R.M. P. (1966). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, 10(3), 516-517. Retrieved from *rasch.org/rmt/rmt103a.htm*.

APPENDIX A:
MCCQE Part I Exam Centres

Alberta	Calgary	University computer lab
	Edmonton	University computer lab
British Columbia	Kelowna	University computer lab
	Prince George	University computer lab
	Vancouver	University computer lab
	Victoria	University computer lab
Manitoba	Winnipeg	University computer lab
New Brunswick	Moncton	University computer lab
Newfoundland	St. John's	University computer lab
Nova Scotia	Halifax	University computer lab
Ontario	Hamilton	University computer lab
	Kingston	University computer lab
	London	University computer lab
	Mississauga	Private lab
	Ottawa	University computer lab
	Sudbury	University computer lab
	Thunder Bay	University computer lab
	Toronto Bay St	Private lab
	Toronto University	University computer lab
Quebec	Chicoutimi	University computer lab
	Montreal I	University computer lab
	Montreal II	University computer lab
	Québec	University computer lab
	Sherbrooke	University computer lab
	Trois-Rivières	University computer lab
Saskatchewan	Saskatoon	University computer lab

APPENDIX B:
MCCQE Part I Statement of Results



MEDICAL COUNCIL OF CANADA / LE CONSEIL MÉDICAL DU CANADA

1021 Thomas Spratt Place
1021, place Thomas Spratt
Ottawa, ON
Canada K1G 5L5
613-521-6012

MEDICAL COUNCIL OF CANADA
QUALIFYING EXAMINATION PART I
STATEMENT OF RESULTS

June 19, 2017

Candidate name:	Xxxxxxx, Xxxxx	**Your final result:**	Pass
MCC candidate code:	xxxxxxxxxxx	**Your total score:**	481
Examination session:	MCCQE Part I Spring 2017	**Score required to pass:**	427

On behalf of the Central Examination Committee, I am writing to inform you of your final result on the Medical Council of Canada Qualifying Examination (MCCQE) Part I that took place during the above-mentioned session.

Your total score, which represents your overall performance, is reported as a scaled score ranging from 50 to 950. Your final result (e.g., pass/fail) is based on your total score relative to the score required to pass. Additional information, including the mean and standard deviation, is available from the MCCQE Part I Scoring web page.

Supplemental feedback on your examination performance is reported to you in a separate document within your physiciansapply.ca account.

Please accept my best wishes for future success.

M. Ian Bowmer, MD CM, FRCPC
Executive Director and Registrar
Medical Council of Canada

mcc.ca
physiciansapply.ca
inscriptionmed.ca

SAMPLE

APPENDIX C:
MCCQE Part I Supplemental Feedback Report

MEDICAL COUNCIL OF CANADA / LE CONSEIL MÉDICAL DU CANADA

1021 Thomas Spratt Place
1021, place Thomas Spratt
Ottawa, ON
Canada K1G 5L5
613-521-6012

SUPPLEMENTAL FEEDBACK REPORT

Candidate name: Xxxxxxx, Xxxxxx
MCC candidate code: xxxxxxxxxx
Your total score: 481
Examination: MCCQE Part I
Examination session: Spring 2017

The purpose of this report is to provide you with supplemental information on your relative strengths and weaknesses, based on your performance across the different domains that were assessed by the test form of the Medical Council of Canada Qualifying Examination (MCCQE) Part II that was administered to you.

Figure 1 displays your performance for the following seven domains: Family Medicine, Medicine, Obstretrics and Gynecology, Pediatrics, Population Health and the Ethical, Legal and Organizational aspects of medicine (PHELO), Psychiatry, Surgery and a Clinical Decision Making component. Each domain is sampled a number of times, with some being measured by a large number of questions and others by a smaller number of questions. Please note that the Family Medicine subscore was generated using responses from subsets of questions from all domains.

To help you better understand your performance, your subscore for each domain is shown along with the mean score of candidates who were first-time takers of the MCCQE Part I in the past year and who passed. The standard error of measurement (SEM) associated with each of your subscores represents the expected variation in your subscore if you were to take this examination again with a different set of questions covering the same or similar domains. Small differences in subscores or overlap between SEMs are indicative that performance in those domains was relatively similar. Likewise, overlap between the SEM for a domain subscore and the mean score of first-time takers who passed, within a given domain, signifies that performance is similar to the mean.

It is important to note that the subscores are based on significantly less data than the total score and that these do not have the same level of precision as the total score. If you have failed the examination and wish to retake it, preparation for all domains is important; otherwise you could improve some subscores and inadvertently lower others.

For more information, please visit the MCCQE Part II Scoring web page.

Report date: 2017-06-19 1/2

SUPPLEMENTAL FEEDBACK REPORT

Figure 1. MCCQE Part I Score Profile