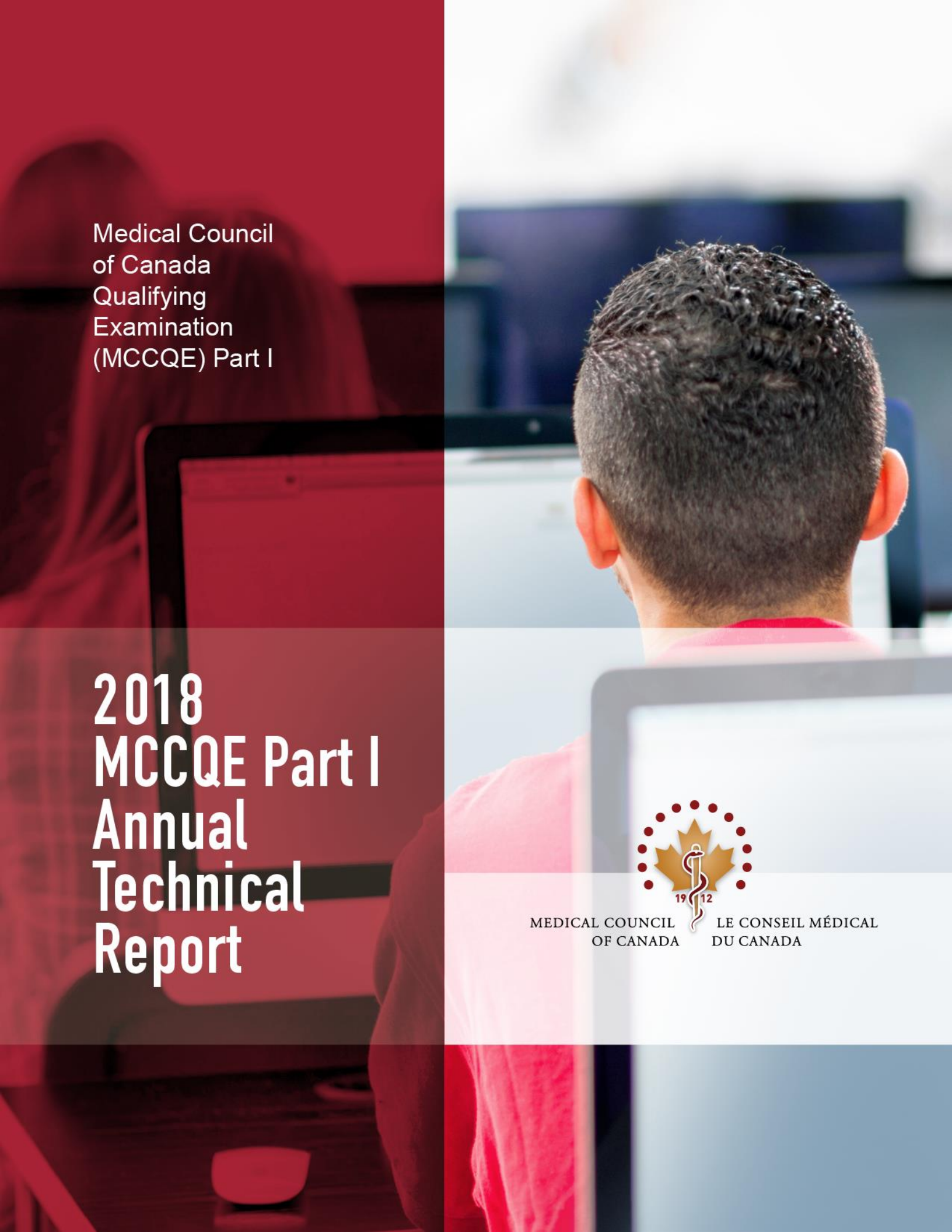
Medical Council of Canada Qualifying Examination (MCCQE) Part I

2018 MCCQE Part I Annual Technical Report



MEDICAL COUNCIL OF CANADA
LE CONSEIL MÉDICAL DU CANADA

Table of Contents

PREFACE **4**

1. OVERVIEW OF THE MCCQE PART I **4**

2. EXAM DEVELOPMENT **5**

2.1 EXAM BLUEPRINT 5

2.2 Exam specifications 8

2.2.1 Content specifications *8*

2.2.2 Psychometric specifications *10*

2.3 Item development 10

2.3.1 Test Committees *11*

2.3.2 Automated item generation *13*

2.3.3 Clinical decision-making items *14*

2.4 Test assembly 15

3. EXAM ADMINISTRATION **18**

3.1 Exam centres 18

3.2 Exam security 18

3.3 Exam preparation 19

3.4 Quality assurance 20

3.5 Release of results 20

4. VALIDITY **21**

4.1 The argument-based approach to validation 21

5. PSYCHOMETRIC ANALYSES **26**

5.1 Item analysis: Classical test theory and item response theory 26

5.2 IRT item calibration 28

5.3 Estimating candidate ability 29

5.4 Scoring 29

5.5 Standard setting and scaling 31

5.6 Score reporting 32

6. EXAM RESULTS **33**

6.1 Candidate cohorts 33

6.2 Overall exam results 33

6.3 Reliability of exam scores and classification decisions 35

6.4 Pass/fail decision accuracy and consistency 37

6.5 Domain subscore profile 37

6.6 Historical pass rates 39

7. REFERENCES **40**

APPENDIX A: MCCQE PART I EXAM CENTRES **42**

APPENDIX B: MCCQE PART I STATEMENT OF RESULTS **43**

APPENDIX C: MCCQE PART I SUPPLEMENTAL INFORMATION **44**

APPENDIX D: INTERNAL STRUCTURE: NEW BLUEPRINT **47**

APPENDIX E: **50**

List of Tables and Figures

Table 1: Blueprint for the MCCQE Part I 7
Table 2: Test Constraints 8
Figure 1. Target test information function 10
Figure 2. Test form representation 15
Figure 3. Automated test assembly procedure 17
Figure 4: Key elements in Kane's argument-based approach to validation: Inferences from observation to decision 22
Table 3: Level of inference – Evaluation/Scoring 23
Table 4: Level of inference – Generalization 24
Table 5: Level of inference – Extrapolation 25
Table 6: Level of inference – Decisions 25
Table 7: Group composition – 2018 33
Table 8: Exam results – spring and fall 2018 34
Figure 5: Total exam score distributions – spring and fall 2018 35
Figure 6. Total exam standard errors of ability – spring 2018 36
Figure 7. Total exam standard errors of ability – fall 2018 36
Table 9: Reliability estimates, standard errors of measurement, decision consistency and decision accuracy indices for each administration of 2018 37
Figure 8: Domain subscore for the spring 2018 38
Figure 9. Domain subscore for the fall 2018 38
Table 10: Spring 2016 to fall 2018 pass rates 39

Preface

This report summarizes the fundamental psychometric characteristics, test development, test publishing, and test administration activities of the Medical Council of Canada Qualifying Examination (MCCQE) Part I. Candidate performance data on the exam in 2018 are also presented. 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 2018 and includes historical data for reference purposes. The report serves as technical documentation and reference materials for members of 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 summative examination that assesses the critical medical knowledge and Clinical Decision-Making (CDM) 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* (Frank, Snell & Sherbino, 2015). Candidates graduating and completing the MCCQE Part I typically enter supervised practice. Aside from the formal accreditation processes of the undergraduate and postgraduate education programs, the MCCQE Part I is the only national standard for medical schools across Canada and is, therefore, administered at the end of medical school.

The MCCQE Part I is a one-day, computer-based test. Candidates are allowed up to four hours in the morning session to complete 210 Multiple-Choice Questions (MCQ), and up to three and a half hours in the afternoon session for the CDM component, which consists of 38 cases with short-menu and short-answer write-in questions. The MCQ portion of the exam is delivered in the morning and the CDM portion is delivered in the afternoon.

The Medical Council of Canada (MCC) undertook a strategic review of its assessment processes with a clear focus on their purposes and objectives, their structure and alignment with the MCC's major stakeholder requirements. The review addressed current trends in medical education, regulation and assessment. The review also considered the role and purpose of the MCC's examinations in meeting the current and future needs of Medical Regulatory Authorities (MRAs), the public and other stakeholders. In addition to focusing on the reassessment and realignment of the MCC's exams, a key recommendation focused on validating and updating the blueprints for both components of the MCC Qualifying Examination (MCCQE).

As part of its commitment to adhere to best practices in medical education and assessment, the MCC undertook a Blueprint project to review and establish an evidence-based approach for identifying the competencies that physicians will be expected to demonstrate and be assessed on at two decision points: (1) entry into residency and (2) entry into independent practice. The purpose is to ensure that critical core competencies, knowledge, skills and behaviours for safe and effective patient care in Canada are being appropriately assessed for the two decision points. The rigorous and consultative process of how the Blueprint was developed can be found *here*.

A new Blueprint for the MCC Qualifying Examinations was approved by Council in 2014 (see section 2.1).

The Central Examination Committee (CEC) is responsible for overseeing the MCCQE Part I including exam blueprint, test specifications and constraints, development of the exam, maintenance of its content, and the approval of results.

2. Exam development

In this section, we describe the exam blueprint, exam specifications, exam format, item development and test development.

2.1 EXAM BLUEPRINT

Exam development begins with the exam blueprint. The exam blueprint for the MCC Qualifying Examinations was approved by Council in 2014. The content specifications for the MCCQE Part I

were approved by the Central Examination Committee in 2016. The blueprint addresses candidates' performance across two broad categories:

- Dimensions of Care, covering the spectrum of medical care;
- Physician Activities, reflecting a physician's scope of practice and behaviours.

There are four domains under each of these two categories. Dimensions of Care reflect the focus of care for the patient, family, community and/or population. Its four assessed domains are:

- **Health Promotion and Illness Prevention**: the process of enabling people to increase control over their health and its determinants, and thereby improve their health. Illness Prevention covers measures not only to prevent the occurrence of illness, such as risk factor reduction, but also to arrest its progress and reduce its consequences once established. This includes but is not limited to screening, periodic health exam, health maintenance, patient education and advocacy, and community and population health.
- **Acute**: brief episode of illness within the time span defined by initial presentation through to transition of care. This dimension includes but is not limited to urgent, emergent and life-threatening conditions, new conditions, and exacerbation of underlying conditions.
- **Chronic**: illness of long duration that includes but is not limited to illnesses with slow progression.
- **Psychosocial Aspects**: presentations rooted in the social and psychological determinants of health and how these can impact well-being or illness. The determinants include but are not limited to life challenges, income, culture, and the impact of the patient's social and physical environment.

Physician Activities reflect the scope of practice and behaviours of a physician practising in Canada and has four domains:

- **Assessment/Diagnosis**: exploration of illness and disease using clinical judgment to gather, interpret and synthesize relevant information that includes but is not limited to history taking, physical examination and investigation.
- **Management**: process that includes but is not limited to generating, planning, organizing safe and effective care in collaboration with patients, families, communities, populations and other professionals (e.g., finding common ground, agreeing on problems and goals of care, time and resource management, roles to arrive at mutual decisions for treatment, working in teams).

- **Communication**: interactions with patients, families, caregivers, other professionals, communities and populations. Elements include but are not limited to relationship development, intra- and inter-professional collaborative care, education, verbal communication (e.g., using patient-centered interviews and active listening), non-verbal and written communication, obtaining informed consent and disclosure of patient safety incidents.
- **Professional Behaviours**: attitudes, knowledge and skills related to clinical and/or medical administrative competence, communication, ethics, as well as societal and legal duties. The wise application of these behaviours demonstrates a commitment to excellence, respect, integrity, empathy, accountability and altruism within the Canadian health-care system. Professional behaviours also include but are not limited to self-awareness, reflection, life-long learning, leadership, scholarly habits and physician health for sustainable practice.

Table 1 displays the new blueprint and associated content specifications (content weightings) for the MCCQE Part I. Both categories, Dimensions of Care and Physician Activities, have four domains, and each domain is assigned a specific content weighting on the exam.

Table 1: Blueprint for the MCCQE Part I

Physician activities	Dimensions of care: Health Promotion & Illness Prevention	Acute	Chronic	Psychosocial Aspects	Row %
Assessment/ Diagnosis					45±5
Management					35±5
Communication					10±5
Professional Behaviours					10±5
Column %	20±5	35±5	30±5	15±5	100

2.2 EXAM SPECIFICATIONS

For the examination to test a broad sampling of topics and populations in medicine as outlined in the blueprint, the MCC has developed content specifications that include certain constraints as well as psychometric specifications. While the exam is divided into two components for delivery purposes – an MCQ component in the morning and a CDM component in the afternoon – content and psychometric specifications are considered at the total test level.

2.2.1 Content specifications

Table 1 contains the content specifications as shown by the content weightings for each of the eight domains.

Table 2 displays the approved test constraints for the MCCQE Part I.

Table 2: Test constraints

CONSTRAINT CATEGORY	DESCRIPTION	CONDITION
Complexity	Multiple morbidities	At least 10%
Age	Neonate, infant/child, adolescent, adult, adult women of childbearing age, and the frail elderly	Sample across the age categories including adult woman of childbearing age and the frail elderly
Gender	Male, female	Balance evenly (minimum of 40% each)
Special populations	Included but not limited to immigrant, LGBT, rural, disabled, and First Nation populations; end of life patients, refugees, inner city poor, the addicted and the homeless	Representative sampling
Setting	Included but not limited to rural or remote settings, long term care institutions and home visits	Representative sampling

The MCQ and CDM components of the MCCQE Part I are described in more detail below.

2.2.1.1 The MCQ component

The MCQ component of the MCCQE Part I consists of 210 items, of which 35 are pilot items that do not count towards the total score. While the pilot items are not scored, they are not identified as pilots within the exam. Each MCQ has an item stem and five options, of which only one is the correct answer. Candidates may select only one option in the

MCQ component of the exam. The maximum time allotted for this component is four hours.

All MCQ questions are presented in a single block. Certain test items will have pictorial material, such as photographs, diagrams, radiograph, electrocardiograms, and graphic or tabulated material.

2.2.1.2 The CDM component

The CDM component of the exam consists of 38 cases, of which eight are pilot cases that do not count towards the total score. While the pilot cases and items are not scored, they are not identified as pilot cases in the exam. Each case includes a case description, followed by one or more items, which assess problem-solving and decision-making skills in the resolution of a clinical case. Candidates may be asked to:

- Elicit clinical information,
- Order diagnostic procedures,
- Make diagnoses, or
- Prescribe therapy.

In total, candidates are presented with 60 to 70 items related to the 38 CDM cases. Items are either in a short-menu or write-in format.

Most items explicitly state how many responses can be selected. Points are not deducted for incorrect answers. However, if a candidate exceeds the maximum number of allowable responses or selects a response that is considered harmful or dangerous to the patient, they will receive a score of zero, even if they have also identified the correct answer. Some items ask candidates to, “select as many as appropriate”. These question types require the candidate to narrow in on the investigation or diagnosis. Selecting too many responses may also result in the candidate receiving a zero, even if the correct answer is part of their answer choice. The maximum time allotted for the CDM component of the exam is three and a half hours.

All cases and questions are presented in a single block. Certain test items will have pictorial material, such as photographs, diagrams, radiograph, electrocardiograms, and graphic or tabulated material.

2.2.2 Psychometric specifications

Psychometric specifications include the desired psychometric properties of the exam, which for the MCCQE Part I includes an overall target Test Information Function (TIF) for each exam form. The target TIF is used to balance multiple forms and to ensure that precision of measurement across the ability scale is highly comparable from one test form to another. Figure 1 displays the target TIF. Test forms are assembled to control maximum information to be within ± 5 per cent of the target.

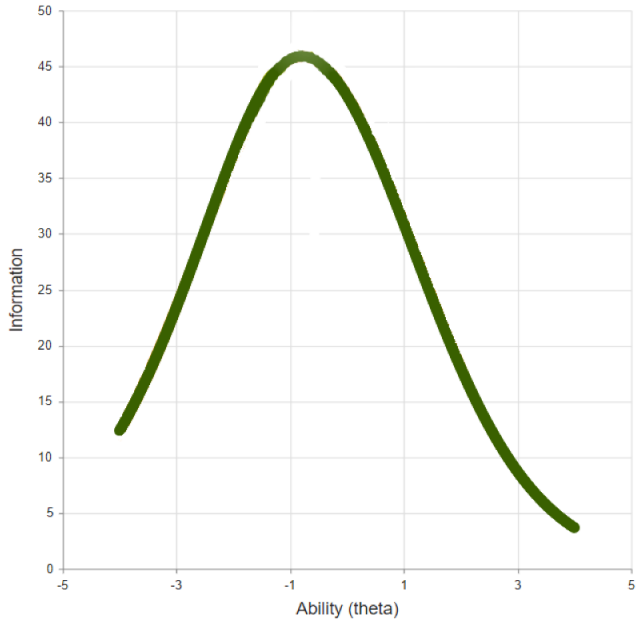


Figure 1. Target test information function

2.3 ITEM DEVELOPMENT

For the MCQ content, six specialty test committees create and approve exam content. For the CDM content, one multi-disciplinary test committee develops exam content. The difference in the CDM Test Committee composition and process is described below in section 2.3.3. MCC's Medical Education Advisor, an expert in medical education and assessment, attends each MCCQE Part I test committee meeting. The Medical Education Advisor trains item writers, educates members on the blueprint and objectives, supports the Test Development Officers (TDOs) in identifying content gap areas, and is a consistent member across committees.

MCCQE Part I content is based primarily on topics that reflect the *MCC Objectives* and align with the approved MCCQE Blueprint. Item writers select a Dimension of Care and a Physician Activity from the Blueprint to write their questions. They also consider test constraints, such as gender, age group, and special populations, during question development as delineated in Table 2.

Each MCQ and CDM Test Committee reviews and approves new content for piloting. New questions are piloted before being used as operational items (active). After the exam administration, candidates' response patterns to pilot items are analyzed. If pilot items meet statistical criteria, they are considered for use in future administrations of the exam. If pilot items do not meet statistical criteria, they are reviewed by test committee members to ensure that the item is defensible. If so, the items are considered for use in future administrations of the exams. If there is an issue detected with an item, it can be discarded or revised and then repiloted.

In the sections that follow, we describe the test committee structure and process we use for developing MCQs and CDMs, the automated item generation process we use to create some MCQs, special considerations for developing CDM items, the process for translating items from English to French and a summary of 2018 item development efforts.

2.3.1 Test Committees

Each test committee is comprised of 8 to 12 Subject Matter Experts (SMEs) from across Canada who have an interest and expertise in the fields of medical education and assessment. Each test committee consists of a minimum of two family physicians. Membership also includes representation from both official language groups (English and French) as content is produced and/or translated in both official languages.

Test Committee membership recommendations can come from TDOs, test committee members, or a member of MCC's Selection Committee. The Selection Committee reviews and approves appointment recommendations at the MCC's Annual General Meeting and formally invites new members to be part of the recommended test committee.

Each test committee meets for two to three days, at least once a year, at the MCC's head office in Ottawa. During these meetings, MCQ and CDM items are written, classified, peer-reviewed and approved by the committee for piloting. There are additional Quality Assurance (QA) processes after the initial committee approval including editorial, which is outlined below.

Committees develop content 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 section 2.3 of the International Test Commission Guidelines on Test Use (2001). These standards and guidelines include QA steps to ensure a fair assessment is delivered to the test takers.

In conjunction with the Chair of each test committee, TDOs guide test committee members in the development of content where identified gaps exist in the exam blueprint, test specifications and constraints. Item development focuses on creating items with a range in level of difficulty and using the most up-to-date medical terminology (for example, compliant with the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders [DSM-5] or newly established guidelines). Committee members focus the development of items content using specific in-practice examples along with anticipating where errors may occur.

After the test committee vets and approves items, the Examination Content Editors ensure the content meets style guidelines, corrects grammar, spelling and punctuation, and conducts fact checking, as required. At times, editors may suggest different words to clarify the meaning of a question. Once the English version of the content is established, a final review of the content is sent to the Multidisciplinary Pilot Approval Committee (MPAC). This committee conducts medical proofing, validates the correct answer, and does a final vetting of the English draft before sending the content to editorial for a substantive edit. Once edited, all content is sent for translation.

Translation of content is outsourced. Since the MCC requires the highest quality of medical translation, all translators go through a screening process to evaluate their qualifications. A comprehensive description of the translation process is summarized in section 2.3.4.

After translation, the MCC engages francophone faculty to ensure that the language in French is inclusive of regional differences in Quebec. The TDOs and Examination Content Editors complete an in-depth comparative read and validation of English and French items. They then engage bilingual test committee members for an out-loud, comparative read of all 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 three-step process by which cognitive models are used to generate items with computer technology (Gierl & Haladyna, 2013):

- **Step 1:** Medical experts identify a content area suitable for item generation. This content is used for the development of a cognitive model.
- **Step 2**: Medical experts create an item model that specifies 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 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. It is important to note that AIG is a tool to augment the development of items rather than replace traditional item development.

By January 2016, using iButler, AIG was launched operationally within all MCQ test committees. The concept was introduced by a half-day training session on AIG, followed by an interactive group exercise on how to create cognitive maps. Lastly, a tutorial was developed to educate members on inputting the data and coding into the iButler software.

In 2017, each test committee meeting was tasked with generating 80-100 items and selecting 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. It was important to note that all items generated through this process were identified as enemies to prevent them from appearing on the same test form.

Overall, the feedback received from committees on the AIG approach to developing MCQs was positive. A generalizability study conducted by the MCC indicated that the variance explained by the modality of item development (AIG vs Traditionally written) is close to 0.0,

suggesting no effect of the modality on exam performance (see Appendix 3). This suggests the interchangeability of items created using different methods. AIG is incorporated as part of regular ongoing activities to supplement traditionally developed items. In 2018, approximately, 25 per cent of pilot MCQs administered were developed using the AIG method.

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, Pediatrics, Population Health, Ethics and Legal Organization of Medicine [PHELO], Psychiatry, Surgery and Family Medicine). The CDM Test Committee has physician representation from both official languages (English and French). Gender diversity and geographic representation from across Canada is also a consideration in the committee membership. Similar to the content development of MCQs, the CDM Test Committee develops content by following professional standards mentioned in section 2.3.1 and rigorous QA processes. Committee members meet twice per year and their mandate is to create, review and classify CDM content based on existing blueprint 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 (three and a half hours), the 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 in terms of item format (short-menu and 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 in clinical scenarios. These scenarios often require critical

decisions to be made during the assessment and management of a given clinical problem. 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 test committee develops the case and questions in accordance with the scenario and the selected MCC Objective. The CDM scoring key reflects the main tasks that candidates must perform, which are identified in the key features. The CDM Test Committee approves all developed cases before they are piloted. As an additional QA step, the six MCQ specialty test committees vet the content and, if necessary, send feedback suggesting revisions to the CDM Test Committee. MPAC also reviews all CDM cases for final medical proofing. Once a case has been piloted and has performed well, the case is banked as an active case ready to be used on a future exam.

Item performance varies and at times, items are flagged for psychometric reasons. All flagged items must be reviewed prior to scoring the exam. Depending on the item, some content will be removed from scoring and must be sent back to committee for review.

2.4 TEST ASSEMBLY

Following item development and piloting, fixed linear test forms are created to meet content specifications, test constraints and psychometric specifications. The number of forms is based on an analysis of operational (active) and field test (pilot) items in the item bank. Due to the number of items per test form and the number of forms, computer software is used in the assembly of the test forms to ensure the construction of equivalent forms, both in content and in difficulty.

As part of test assembly, we take into account linking. Scores from different test forms are statistically linked through common items referred to as *anchor items* (see Figure 2). These items are shared between adjacent test forms, and they ensure score comparability across test forms.

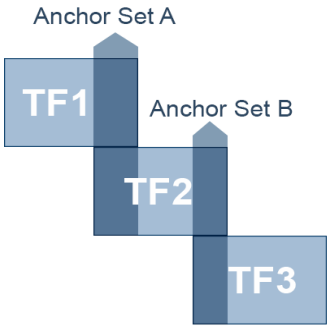


Figure 2. Test form representation

Anchor items are assembled as a set of MCQs called *anchor sets*. There are no CDM anchor sets currently. Most test forms contain two anchor sets for linking purposes, except for the first and last test form. Anchor items are selected using the content specifications to be a smaller representation of a complete exam in terms of both content and psychometric specifications and content constraints.

TDOs collaborate with psychometricians and physicians in the assembly of multiple test forms to ensure candidates receive a broad representation of content in their test-taking experience in line with the content specifications, test constraints and psychometric specifications. Other guidelines used in the assembly of the tests include ensuring the appropriate representation of topics of medicine, confirmation that items refrain from providing answers to other test questions and that item enemies (items of similar content) are tracked to avoid appearing on the same test form, and tracking AIG items across the test forms.

The TDOs and psychometricians work closely to ensure the test forms, in their entirety, are reviewed and approved by SMEs. Once MCC staff has vetted the forms to ensure they meet the exam specifications, two different committees of SMEs convene once per year to review and approve the test forms. The first committee is the Anchor Set Approval Committee (ASAC) and the second is the Test Form Approval Committee (TFAC).

Both the ASAC and TFAC follow a similar, thorough process to approve the test forms using the MCC's Test Form Management (TFM) system. The process for form approval is:

1. The Psychometrics and Assessment Services (PAS) staff assemble test forms according to the exam specifications.
2. The Evaluation Bureau's (EB) TDOs approve the forms, exchanging any items that overlap in content or may pose as item enemies not yet tagged in MOC5. TDOs also identify any content that may be medically inaccurate (e.g., guideline changes).
3. The ASAC approves the MCQ anchor sets first, as they establish the linking scale that connects all forms to ensure a comparable level of difficulty and precision. Once approved, the Anchor sets are considered "locked" (i.e., they cannot be replaced during the approval of an entire form).
4. The TFAC then reviews the remaining items on each test form and approves all the forms in their entirety.
5. Pilot forms are then also approved by TFAC.

6. A final review by PAS and the TDO ensures the content specifications and constraints have been respected and the psychometric parameters are maintained in the final approved forms.

The MCCQE Part I has evolved from a semi-adaptive exam, where questions candidates saw depended on their responses to previous items, to fixed examination forms where a pre-selected set of items is included in each form. MCC has developed automated methods for assembling test forms through constrained optimization that can most efficiently support the construction of multiple parallel test forms. After forms are assembled, they are reviewed and approved by the MCC's MCCQE Part I team (which includes item and test development experts and psychometricians) and two independent committees of physicians. Automated Test Assembly (ATA) was used to assemble all MCCQE Part I test forms. Test forms were assembled to meet a series of content specifications, as described in section 2.2, and to be as similar as possible, both in content and in difficulty. Figure 3 depicts the logic implemented to automatically assemble a number of test forms. Common items between test forms are required to establish a common scale for item parameter estimates obtained from different test forms. The result is that scores from different test forms can be compared as they share a common scale.

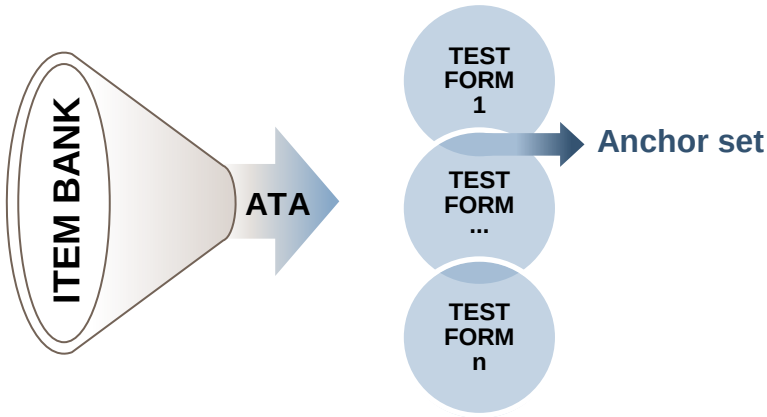


Figure 3. Automated test assembly procedure

The TIF for each of the test forms was inspected. The maximum information for each form were within ± 5 per cent of the target value. TIF can be used to observe how much information an item contributes and to what portion of the scale score range. It also shows the degree of precision at different values of candidates' ability, as information is defined as the reciprocal of the precision with which a parameter could be estimated.

3. Exam Administration

3.1 EXAM CENTRES

In 2018, the MCCQE Part I was offered during two test windows in April/May and October/November. The test window occurs over a two-to-four-week period, at 26 sites, in both university computer labs and private testing centres across Canada.

MCC staff delivers and monitors the exam through the QEI.net system developed by the MCC Information Technology (IT) department in 2001.

During the exam, site coordinators, who administer the exam at the faculties of medicine or private testing centres, are required to call in to MCC staff each morning, using a dedicated phone number, to access security permissions. These permissions allow them to log into the secure exam system. Each site coordinator has a personal identification code that they must enter along with the candidate's code and Personal Identification Number (PIN) for the exam to launch. Site coordinators work directly with MCC staff to address technical permissions, security issues, technological issues, and emergency situations.

The number of days a test 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 can write the exam 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 through site visits when new sites are opened or when there is a new site coordinator. The MCC also conducts yearly training with all site administrators to communicate enhancements to MCC protocols and reinforce security measures. In addition to test security measures at the test sites and a team that monitors exam activities throughout the examination session, EB staff monitors online study forums for candidate activity around sharing of exam content before, during and after the administration.

Candidates taking a MCC examination have legal and professional responsibilities. The MCC also has a responsibility to candidates and to Canadians to ensure the integrity of its examinations. In 2018, the MCC introduced, as part of its registration and exam day process, an Exam Test Security video (https://mcc.ca/news/mcc-launches-new-test-security-video/). All candidates need to agree to the terms and conditions, which state that they have understood the rules and regulations around test security. The creation of the video was in response to increased content breaches and a pattern from candidates that they were unaware that sharing exam content was in violation of their terms and conditions.

If a candidate appears to be giving or receiving information during the exam, the site administrator can change their location in the exam room or immediately terminate their exam. The site administrator is required to produce a full report of all such occurrences to the MCC. All MCCQE Part I materials, including the content and questions comprising the MCCQE Part I, are protected by copyright and are to be kept confidential. Candidates are permitted to use the MCCQE Part I materials solely for the purpose of completing the MCCQE Part I and must not disseminate, reproduce, share or reveal to others the exam materials and content, in whole or in part, at any time in any way, even after the exam ends. Comparing exam content and question themes with colleagues, sharing content with future exam candidates and posting content online are considered breaches of confidentiality. Any breach of the MCCQE Part I Terms and Conditions is considered irregular behaviour for which the MCC or CEC may take appropriate action, in accordance with the MCCQE Part I Terms and Conditions candidates accepted at time of application. In the past, the CEC has issued a Denied Standing to a candidate, due to irregular behaviour, and a barring from taking future MCC examinations for a period of time.

3.3 EXAM PREPARATION

Online preparatory materials are available to assist candidates prepare for the MCCQE Part I. These resources include the exam platform demonstration videos, sample questions (MCQ &

CDM), instructional videos (CDM tips, online demo, etc.), a list of resources by medical specialty area, and the MCC Objectives. All candidates have access to these materials through the MCC's website (*https://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.

3.4 QUALITY ASSURANCE

After each exam administration, MCC's database is updated with two basic data sets, 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 items. An initial round of QA 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 EB. Once it is determined that the data meets the established QA requirements, scoring and calibration are performed by PAS.

3.5 RELEASE OF RESULTS

Examination results are confirmed by the CEC. Approximately seven weeks after the last day of the examination session, the CEC meets to review performance on the exam, address administrative issues, rule on special candidate cases, and approve exam results.

The MCC releases candidates' final results (e.g. pass/fail decision) and total score through their physiciansapply.ca account. Shortly thereafter, candidates have access to their Statement of Results (Appendix B), the official results document, and the Supplemental Information Report (Appendix C) that provides them with information on their strengths and weaknesses by the domains in the blueprint.

4. Validity

It is generally accepted that tests are not inherently valid or invalid but that validity should be viewed as a process of gathering evidence that supports the intended uses/interpretations of test scores (AERA, APA, & NCME, 2014). Michael T. Kane (1990, 2013a, 2013b) has proposed an argument-based approach to validation that involves a process of gathering evidence to support score interpretations by establishing arguments that can be backed by theory, empirical research or common sense (Kane, 1990).

4.1 THE ARGUMENT-BASED APPROACH TO VALIDATION

According to Kane (2013b), the validity of a proposed interpretation and use depends on the plausibility of the claims being made, and validation involves the evaluation of these claims. Any claim that certain statements about score use or interpretations being valid must be justified. Justification takes on the form of arguments. “Proposed interpretations and uses are valid to the extent that the reasoning involved in the interpretation is sound, reasonable, and plausible, that is, valid” (Kane, 1990).

For the MCCQE Part I, this entails gathering evidence to support the intended uses/interpretations of the examination, namely that scores and pass/fail decisions can be used to make valid decisions regarding the level of competence of a graduating student entering supervised practice. Validity considerations have been incorporated into exam design, exam specifications, item development, exam assembly, psychometric quality, exam administration and results reporting.

In Kane’s approach, validating the interpretive arguments involves four inferences:

1. **Evaluation/Scoring**: Assigning scores to performance
2. **Generalisation**: From statements about observed performance to statements about expected performance over a universe of possible performances
3. **Extrapolation**: Statements are extended to the expected performance over the domain
4. **Decisions/Implication**: Performance can also be used to make decisions about an examinee’s future

Figure 4 depicts Kane's framework for an argument-based approach to validation. His approach begins with an assessment of the Scoring of a single observation (e.g., responses to exam items), to using the observed scores to generate an overall test score representing performance in the test setting (Generalisation), to drawing an inference regarding what the test score might imply for real life performance (Extrapolation), and finally to interpreting this information and making a decision (Implications).

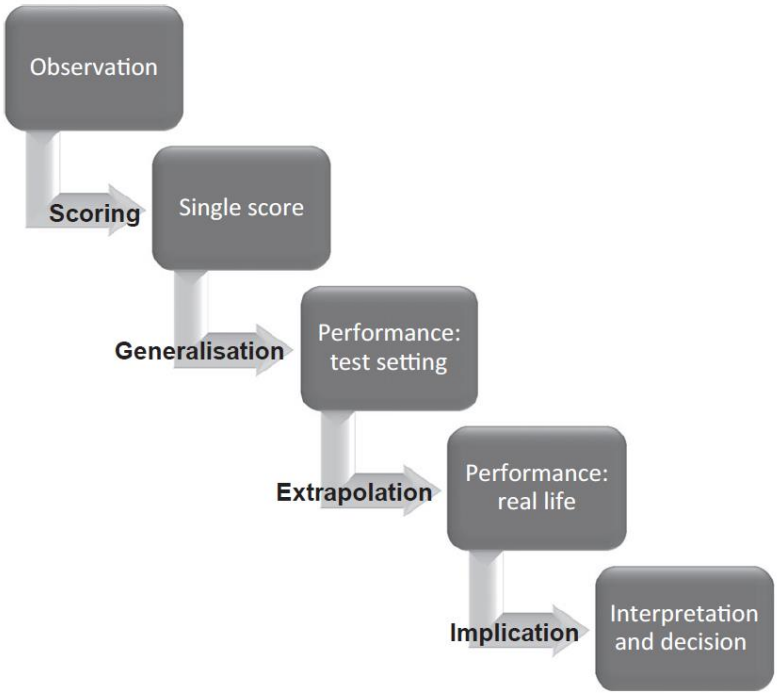


Figure 4: Key elements in Kane's argument-based approach to validation: Inferences from observation to decision (Source: Cook, 2015, page 564)

In Tables 3 to 6, we provide evidence for the four levels of inference of Kane's argument-based approach to validation. In each of these tables, we present information about the Source of Evidence (content expertise, test content, internal structure, etc.), Datum (data used to support the claim), Warrant (logical statements that serve as bridges between the claim and the data), and Backing (additional justification for the warrant).

Table 3: Level of inference – Evaluation/Scoring

Sources of evidence	Datum	Warrant	Backing
Based on content expertise	Documentation, meeting notes, training slides	Items are developed to reflect relevant medical ability	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, items are developed based on content specifications and test constraints defined by the CEC members who ensure that the 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.
Based on content expertise	Documentation, meeting notes, training slides	Proper training is offered for test developers	Various test committees are involved in developing test items. Regular content development workshops are conducted to train test committee members to develop items that reflect the knowledge and skills emphasized in the content specifications 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 items are reviewed, edited and finalized by test committee members, TDOs, editors, and translators.
Based on content expertise	Documentation, meeting notes, training slides	Construct-irrelevant variance is minimized during item development	During development, items are reviewed by SMEs and TDOs to ensure they meet the content 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.
Based on test content	Item responses and scoring rules (MCQs/CDMs)	The answer keys are the correct answers	Expectation is that item-total correlations for correct answers are positive and are negative for distractors; items not meeting this expectation are identified and provided to TDOs for content review before final calibration/test scoring.
Evidence of precision/ Based on internal structure	Write-ins item responses	Markers are marking write-in responses consistently within an exam session	Each item is marked independently by two physician markers and when discrepancies are detected, the issue is resolved by a third marker. 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.

Table 4: Level of inference – Generalization

Sources of evidence	Datum	Warrant	Backing
Evidence of precision	Item and test scores	The reported scores attain the level of decision accuracy and decision consistency meets the target values	The decision consistency estimate and the decision accuracy estimate for the spring administration were 0.90 and 0.93, respectively, which indicates reliable and valid pass/fail decisions. Values were slightly below the target values in fall session given the composition of population taking this session (mostly, int'l medical graduates [IMG]). Detailed information can be found in section 6.3 of this report.
Evidence of precision	Item and test scores	The reported scores attain the level of precision required for a high-stakes exam; total score reliability estimates are above the target values.	Person [test] reliability estimate in spring was 0.88 and in fall 0.85, indicating adequate level of reliability of test scores, given the characteristics of the population of examinees (i.e., high achievers).
Based on test content	Blueprint classification	Test forms are comparable in content	ATA was used to assemble a number of fixed linear test forms, all of which met content specifications and test constraints, as described in section 2.
Based on test internal structure	Item parameters	Test forms are comparable in level of difficulty	During ATA, test forms were assembled to also be as similar in difficulty as possible. TIF for each of the test forms were inspected and results support the parallelism among the different test forms.
Based on test internal structure	Correlation among domains and total score	Blueprint domains are highly correlated with total score	All domains were found to be significantly, positively correlated with one another (see Appendix D). The highest correlation was found with the Total Score. This suggests that the MCCQE Part I seems to measure an essentially single dominant underlying construct (i.e., basic medical knowledge and clinical skills that the MCCQE Part I is designed to measure). Furthermore, this provides preliminary evidence to support the assumption of unidimensionality underlying the use of the Rasch model used to assemble and score the exam.
Based on Generalizability study (G-study)	Item responses and test taker information	Items are performing comparably for Francophones and Anglophones	G-Study results indicate that the variance explained by Language in which the exam was taken is close to 0.0, suggesting no effect of Language on exam performance. Results are presented in Appendix E.
Based on G-study	Item responses and test taker information	Items are performing comparably for male and female	G-Study results indicate that the variance explained by the candidate's gender is close to 0.0, suggesting no effect of gender on exam performance. Results are presented in Appendix E.

Table 5: Level of inference – Extrapolation

Sources of evidence	Datum	Warrant	Backing
Evidence of relationship with other exams	MCCQE Part I test scores / Medical Council of Canada Evaluating Examination (MCCEE) test scores	The correlation between the MCCQE Part I and MCCEE scores provides some evidence of convergent validity	The relationships between scores on the MCCQE Part I and the MCCEE were investigated. A significant correlation (r=.70, p<.0001) was obtained between the exams based on a sample of 447 candidates whose scores on both exams were matched using data from the April 2018 administration of the MCCQE Part I and the five sessions of the MCCEE of 2017.
Evidence of precision	Item and test scores	The correlation between the MCCQE Part I and NAC exams provide some evidence of convergent validity	The relationships between scores on the MCCQE Part I and the NAC Examination were also investigated. The NAC Examination uses an Objective Structured Clinical Examination (OSCE) format to assess the readiness of an IMG for entry into a Canadian residency program. A significant correlation (r=.55, p<.0001) was obtained between scores on the MCCQE Part I and the NAC Exam based on a sample of 87 candidates whose scores on both exams were matched using data from spring 2018. The correlation is strong enough to provide some evidence of convergent validity between the two MCC exams, but not too high to indicate redundancy as the two exams are assessing different aspects of clinical knowledge and skills. Caution advised in interpreting this result due to low number of candidates taking both exams (N=87).

Table 6: Level of inference – Decisions

Sources of evidence	Datum	Warrant	Backing
Based on standard setting	MCCQE Part I test scores and pass/fail status; Subject Matter Expertise	Those who pass the MCCQE Part I are competent enough to practise safely and efficiently	The cut score is reflective of a point on the proficiency scale that represents the minimum standard. After a comprehensive standard-setting procedure with 22 panelists, the MCC's CEC endorsed a pass score of 226 on a scale of 100 to 400 as a defensible standard to apply starting with the April 2018 administration. Sources of validity evidence that the MCCQE Part I meets best practices when setting new pass scores are: careful selection of panelists; careful training of panelists, standard-setting methodology followed best practice (Bookmark and Hofstee methods); and feedback of the panelists post standard-setting exercise. Internal evidence included the consistency of the panelists and convergence of results. Two subpanels arrived at a similar pass score independently at 95% confidence intervals constructed using Standard Error of Judgment (SEJ). SEJ indicates the variability that would be expected if the same judging process was repeated by many different panels of similar composition. More information on the Standard-Setting procedure can be found *here*.

5. Psychometric analyses

In 2018, the MCCQE Part I was offered twice, in April/May (spring) and October/November (fall), during two- to three-week testing windows at both university computer labs and private testing centres across Canada. In this section, we describe the psychometric analyses completed following the spring exam administration. We conduct item analyses, followed by item calibration, estimation of candidates' ability, scoring, standard setting and scaling, and finally, score reporting. After item calibration in the spring, we have pre-calibrated forms that are used for the fall session.

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 (CTT) and Item Response Theory (IRT), 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.

CTT and IRT 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 score correlation is computed to reflect item discrimination. 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). Furthermore, items with near zero option endorsement (for example, too few candidates who obtain a particular score or choose a particular distractor) are also flagged for content review.

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. That is, 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 more able candidates). These items are flagged as potentially flawed and subject to content review. Additionally, CDM write-in items that display less than 90 per cent agreement between markers or have a weighted kappa coefficient of less than 0.61 are also flagged for review. The kappa coefficient reflects the agreement between markers above and beyond chance agreement (Cohen, 1979), as it is expected that scores assigned by two markers would yield highly comparable results.

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
- High percentage of omits: >5 per cent
- Low correlation value for the correct answer: <0.05
- High correlation value for distractor: >0.05 and N>10
- Top 20 per cent performers chose distractor more often than correct answer
- Item mean square outfit < 0.5
- Item mean square outfit > 2.0.
- Low category score frequency N <10
- Disordered Threshold (write-in only)
- Average ability not increasing (write-in only)
- Percent Agreement < 0.90 (write-in only)
- Weighted Kappa < 0.61 (write-in only)

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 are removed from further analyses with the review committee's approval. Following the IIA in spring 2018 and after consultation with content experts, 194 MCQs and 7 CDM items were not included in the final scoring. The fall administration is processed using the same item difficulty estimates from spring and the same poor performing items from the spring are removed in the fall session so that scores are on the same scale and thus comparable.

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 2018 MCCQE Part I, items were freely estimated using data from Canadian Medical Graduates (CMG) first-time test takers. First, the parameters for the all active/operational

items were estimated to identify potential 'poor performing' items. Through this step, items that did not satisfy the statistical criteria outlined in Section 5.1 were flagged and reviewed by SMEs. The decision to be made was to retain or remove those items from scoring. After the TDO makes arrangements with the SMEs to review all flagged items (in Step 1) and provides decision on which items to remove from scoring and calibration, items are recalibrated excluding those items. A final set of calibrated items are then ready to use in estimating candidates' abilities.

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 items' difficulty. 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. When the estimation procedure has reached an acceptable level of convergence, all pre-specified output has been produced, including the file containing the persons abilities.

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 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. Sometimes, CDM items can also be dichotomously scored. 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).

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 regard to the specialty areas being assessed whereas a candidate who obtains a score of +3.00 would demonstrate strong knowledge.

5.5 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 summer of 2018, the MCC completed a rigorous standard-setting exercise1 based on expert judgments from a panel of 22 physicians representing faculties of medicine from across the country, different specialties and years of experience supervising students and residents. The Bookmark Method, a successfully employed and defended method used by large-scale exam programs, was used to help panelists suggest a new pass score for the exam. The recommended pass 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. The CEC approved the recommended pass score.

In the spring 2018 MCCQE Part I, a new pass score was applied to reflect this minimally-acceptable level of performance. The value representing this standard was established at 0.682 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 250 and a standard deviation of 30 based on all first-time candidates in spring 2018. On that scale, the pass score is equivalent to 226 for the MCCQE Part I.

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;

[1] *mcc.ca/media/MCCQE-Part-I-Standard-setting-report-2018.pdf*

$$\text{And} \quad X_{\text{i}} = \text{a candidate's Rasch ability score}$$

In the spring of 2018, when the scale was first established, the slope and intercept were established to be 58.46300753 and 185.7324343, 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.

5.6 SCORE REPORTING

Approximately seven weeks after the last day of the exam session, the MCC issues a Statement of Results (SOR) and a Supplemental Information Report (SIR) to each candidate through their physiciansapply.ca account. Samples of the SOR and SIR can be found in Appendix B and C, respectively. The SOR includes the candidate's final result and total score as well as the score required to pass the exam. Additional information about subscores and comparative information is provided in the SIR, 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 purposes.

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 2018 is summarized in this section. When applicable, historical data from previous years are included for reference.

6.1 CANDIDATE COHORTS

In 2018, the MCCQE Part I was administered in a three-week window (April 16 to May 9) in the spring and in a one and a half-week window (October 29 to November 07) in the fall. A total of 5,408 candidates challenged the exam across the 26 testing sites. Of the total number of candidates who took the examination in 2018, one candidate received a Denied Standing and two candidates were removed from the exam statistics figures and tables pending committee decision in the new year. Table 7 summarizes the distribution of candidates across groups defined by their country of graduation and whether they were a first-time or repeat test taker of the MCCQE Part I.

Table 7: Group composition – 2018

			Fall 2018		Total	
Group	N	%	N	%	N	%[1]
CMG first-time test takers	2810	67.4	13	1.1	2823	52.2
CMG repeat test takers	64	1.5	114	9.2	178	3.3
IMG first-time test takers	733	17.6	681	54.9	1414	26.2
IMG repeat test takers	560	13.4	433	34.9	993	18.4
TOTAL	**4167**		**1241**		**5408**	

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

6.2 OVERALL EXAM RESULTS

Table 8 summarizes pass rates for the 2018 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 100 to 400; the pass score is 226. This table does not include the one candidate who received a Denied Standing or the two candidates who are awaiting a committee decision.

Table 8: Exam results – spring and fall 2018

		Exam Results		
		Spring 2018	Fall 2018	Total
CMG First-time Test Takers	N	2810	13	2823
	M	262	244	262
	SD	22.4	17.5	22.4
	Min.	189	201	189
	Max.	344	265	344
	Pass Rate (%)	95	85	95
CMG Repeat Test Takers	N	64	114	178
	M	231	234	233
	SD	18.5	16.1	17.0
	Min.	192	189	189
	Max.	277	283	283
	Pass Rate (%)	63	70	67
IMG First-time Test Takers	N	733	680	1413
	M	236	232	234
	SD	27.6	27.9	27.8
	Min.	100	141	100
	Max.	315	316	316
	Pass Rate (%)	65	59	62
IMG Repeat Test Takers	N	559	432	991
	M	210	211	211
	SD	20.7	19.9	20.4
	Min.	143	134	134
	Max.	268	265	268
	Pass Rate (%)	23	24	24
All Candidates	N	4166	1239	5405
	M	250	225	244
	SD	29.8	26.3	30.9
	Min.	100	134	100
	Max.	344	316	344
	Pass Rate (%)	80	48	73

Figure 5 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.

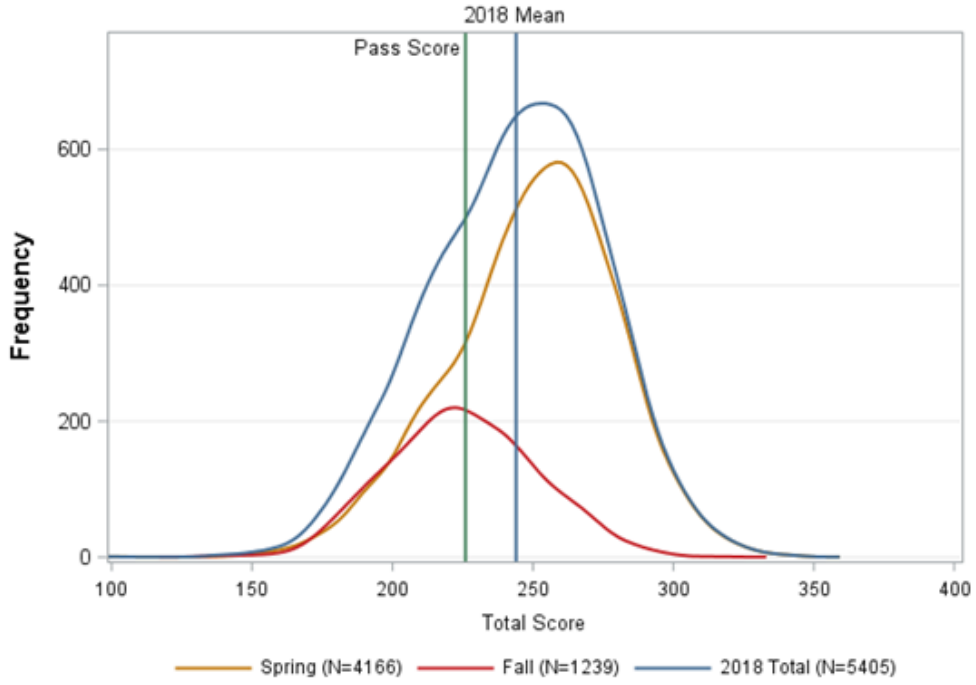


Figure 5: Total exam score distributions – spring and fall 2018

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 6 and 7 display scatter plots of SE values along the reported score scale for the spring and fall 2018 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 lower near the pass score,

which indicates highest precision of ability estimates, thus supporting more accurate and consistent pass/fail decisions.

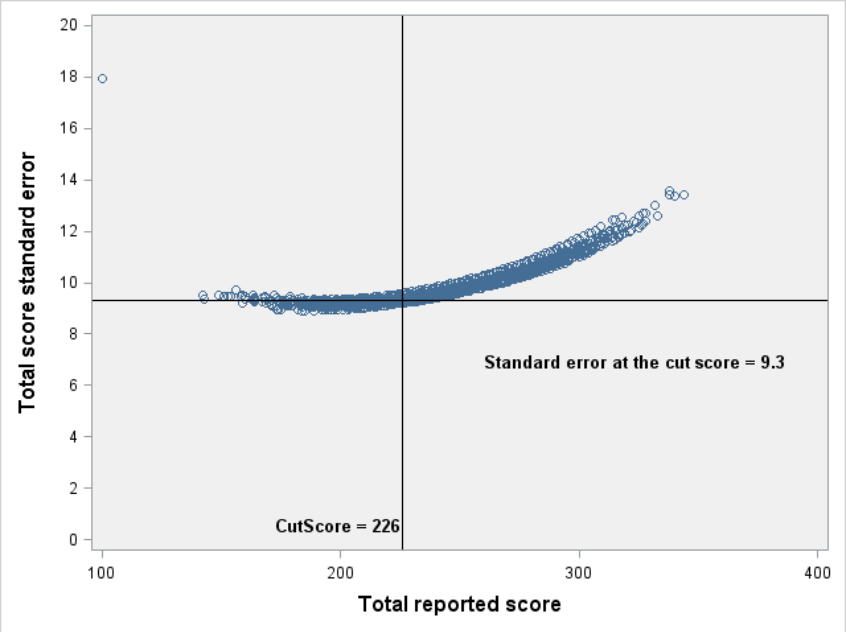


Figure 6. Total exam standard errors of ability – spring 2018

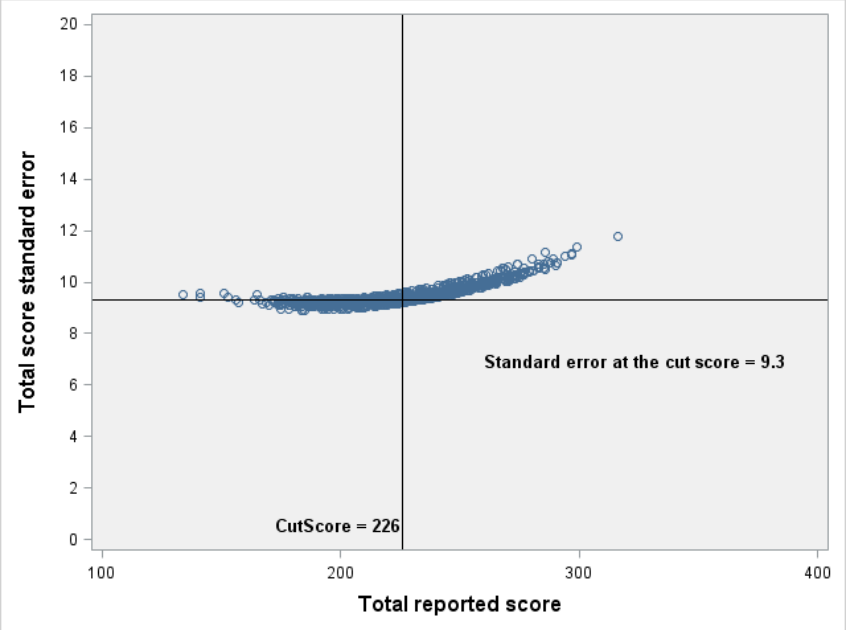


Figure 7. Total exam standard errors of ability – fall 2018

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. 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 2018 indicate reliable and valid pass/fail decisions based on MCCQE Part I scores. Table 9 is based on data from 4166 candidates in the spring session and 1241[2] candidates in the fall session.

Table 9: Reliability estimates, standard errors of measurement, decision consistency and decision accuracy indices for each administration of 2018

	Spring	Fall
Reliability estimate[1]	0.88	0.85
Average SEM (total score)	9.9	9.5
Decision consistency	**0.90**	**0.83**
False positive	0.05	0.09
False negative	0.05	0.09
Decision accuracy	**0.93**	**0.88**
False positive	0.03	0.06
False negative	0.04	0.06

[1] *Person (test) reliability from the Rasch model.*

6.5 DOMAIN SUBSCORE PROFILE

The purpose of the domain subscore profile is to provide diagnostic information to candidates by highlighting their relative strengths and weaknesses. The SIR 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 2018 administrations. The range of domain subscores is presented graphically in Figures 8 and 9. 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

[2] Does not include one candidate who received a Denied Standing.

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).

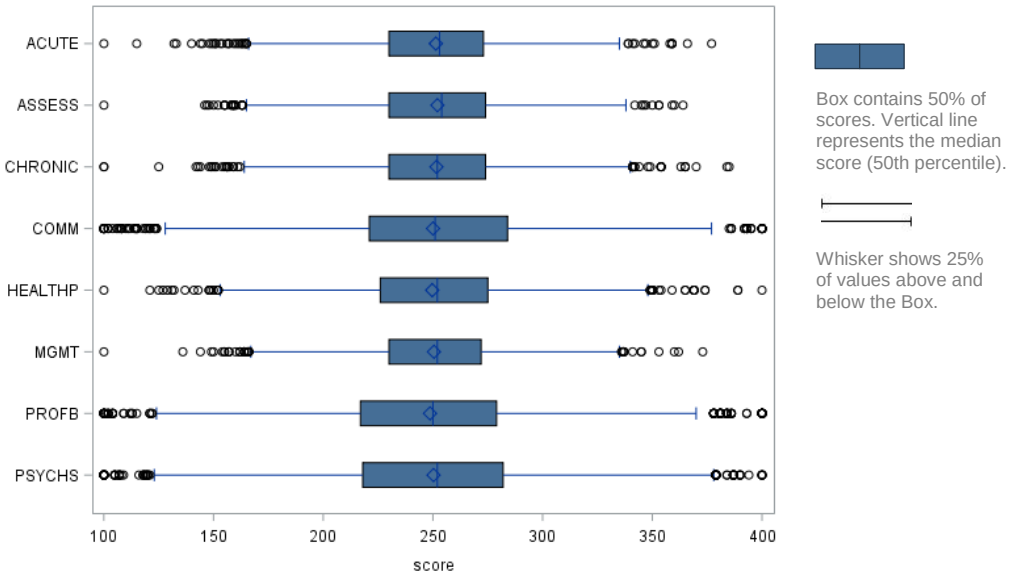


Figure 8: Domain subscore for the spring 2018

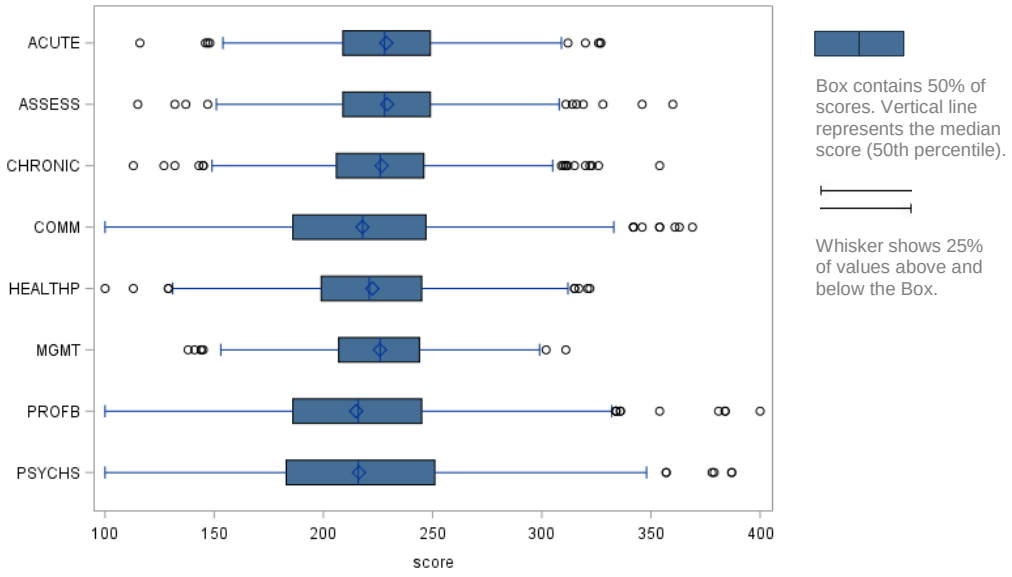


Figure 9. Domain subscore for the fall 2018

6.6 HISTORICAL PASS RATES

Historical pass rates are presented in this section. Table 10 shows the pass rates for 2016 to 2018 by group.

Table 10: Spring 2016 to fall 2018 pass rates

	2016		2017		2018	
	N	Pass rate	N	Pass rate	N	Pass rate
CMG first-time test takers	2831	97	2802	95	2823	95
CMG repeat takers	171	69	156	63	178	67
IMG first-time test takers	1704	58	1677	62	1413	62
IMG repeat takers	1210	29	1264	29	991	24
TOTAL	**5916**	**71**	**5899**	**71**	**5405**	**73**

7. 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.

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*.

Cook D. A., Brydges R., Ginsburg S., Hatala R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ., 49(6):560-75. doi: 10.1111/medu.12678

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*.

Frank JR, Snell L, Sherbino J, editors. CanMEDS 2015 Physician Competency Framework. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015.

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.

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

Kane, M. (1990). An Argument-based Approach to Validation. Iowa City, Iowa: American Coll. Testing Program.

Kane, M. (2013a). The argument-based approach to validation. School Psychology Review, 42(4), 448-457.

Kane, M. (2013b). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1-73.

Linacre, J. M. (2015). Winsteps (Version 3.91.0) [Computer software]. Retrieved from http://www.winsteps.com

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

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*.*

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.

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.

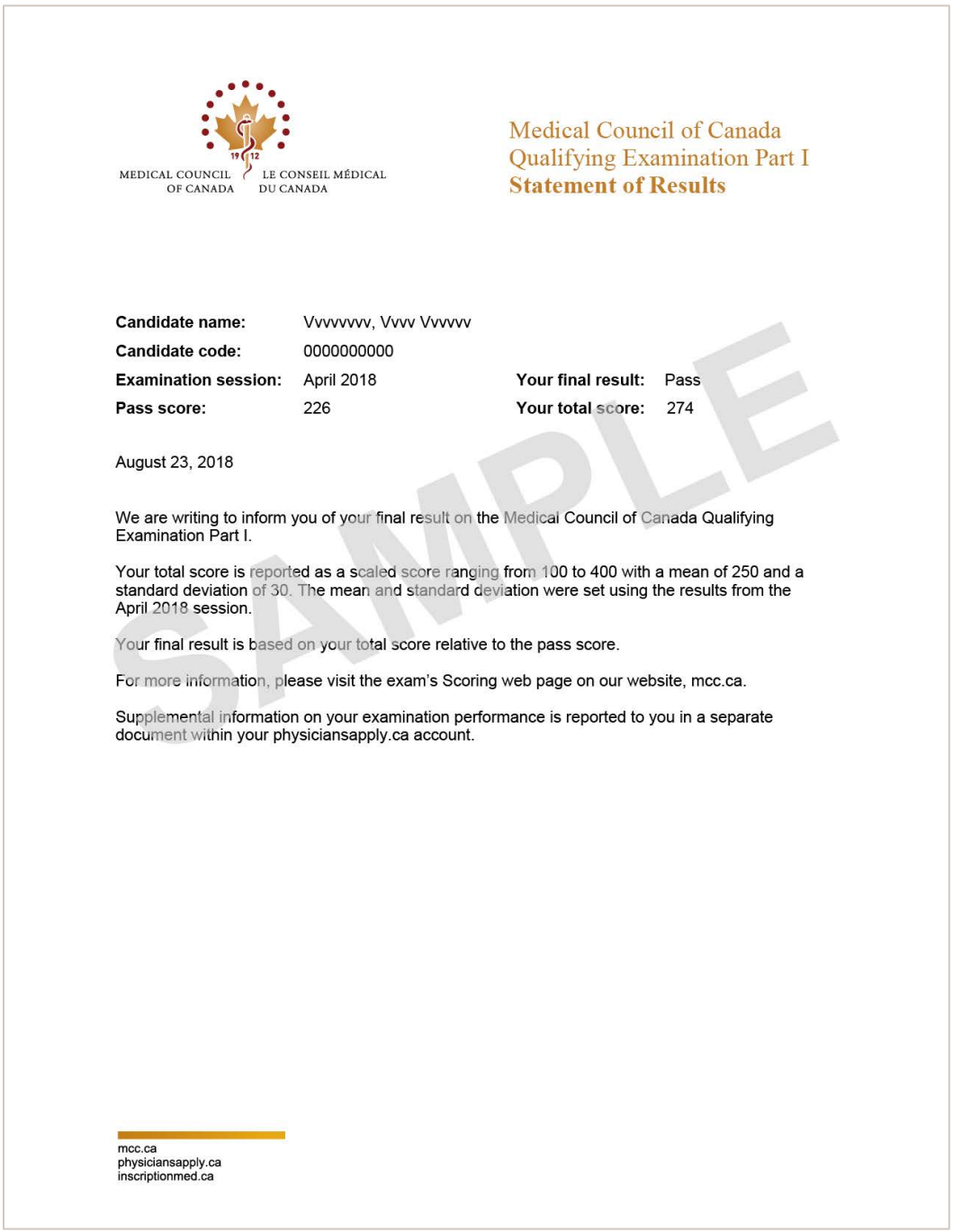
Muchinsky P.M. (1996) The correction for attenuation. Educational & Psychological Measurement 56:1, 63-75.

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.

APPENDIX A:
MCCQE Part I Exam Centres

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

APPENDIX B:
MCCQE Part I Statement of Results



MEDICAL COUNCIL OF CANADA | LE CONSEIL MÉDICAL DU CANADA

Medical Council of Canada
Qualifying Examination Part I
Statement of Results

Candidate name:	Vvvvvvvv, Vvvv Vvvvvv		
Candidate code:	0000000000		
Examination session:	April 2018	**Your final result:**	Pass
Pass score:	226	**Your total score:**	274

August 23, 2018

We are writing to inform you of your final result on the Medical Council of Canada Qualifying Examination Part I.

Your total score is reported as a scaled score ranging from 100 to 400 with a mean of 250 and a standard deviation of 30. The mean and standard deviation were set using the results from the April 2018 session.

Your final result is based on your total score relative to the pass score.

For more information, please visit the exam's Scoring web page on our website, mcc.ca.

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

SAMPLE

mcc.ca
physiciansapply.ca
inscriptionmed.ca

APPENDIX C:
MCCQE Part I Supplemental Information Report

MEDICAL COUNCIL OF CANADA | LE CONSEIL MÉDICAL DU CANADA

Medical Council of Canada
Qualifying Examination Part I
Supplemental Information Report

Candidate name:	Vvvv, Vvvvvv Vvvv		
Candidate code:	0000000000	**Your final result:**	Pass
Examination session:	April 2018	**Your total score:**	274

This report provides you with supplemental information on your performance on the Medical Council of Canada Qualifying Examination (MCCQE) Part I.

The MCCQE Part I 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 exam assessed your performance across two broad categories with each exam question classified on both categories:

- Dimensions of care, covering the spectrum of medical care;
- Physician activities, reflecting a physician's scope of practice.

Each category has four domains:

Dimensions of Care	Physician Activities
Health Promotion and Illness Prevention	Assessment and Diagnosis
Acute Care	Management
Chronic Care	Communication
Psychosocial Aspects	Professional Behaviours

See p. 3 of this report for the domain definitions.

Figure 1 displays your performance in each domain under Dimensions of Care. Figure 2 displays your performance in each domain under Physician Activities.

In both figures, we provide your subscores along with the mean subscore of first-time takers who passed the same exam in spring 2018 when the reporting scale and pass score were established.

Each domain is assigned a weighting on the exam. We present the content weights, expressed as percentages, in the grids shown on page 3.

We also provide the standard error of measurement (SEM) for each of your subscores. It represents the expected variation in your subscore if you were to take this exam again with a different set of questions covering the same domains.

Small differences in subscores or overlap between SEMs indicate that performance in those domains was somewhat similar. Overlap between the SEM and the mean score of first-time takers who passed signifies that performance is similar to the mean.

Subscores are based on less data than the total score and have less precision.

For more information, please visit the exam's Scoring web page on our website mcc.ca.

Figure 1: Dimensions of Care

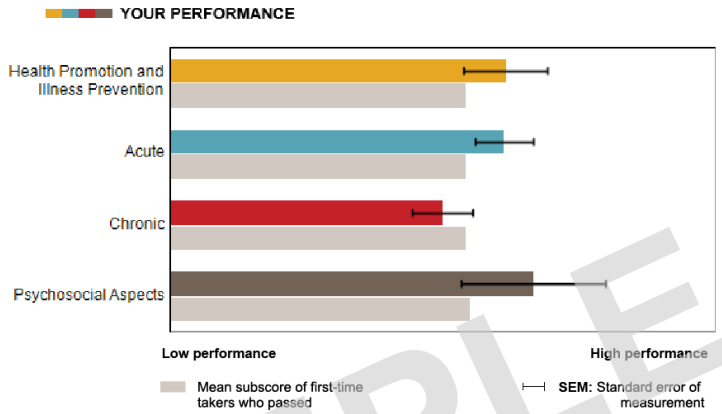
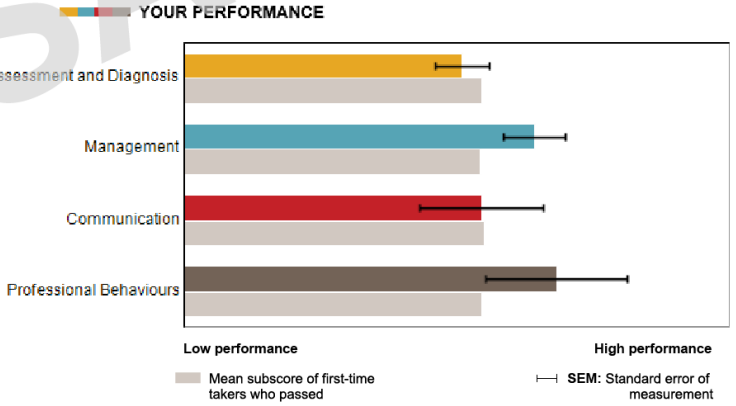
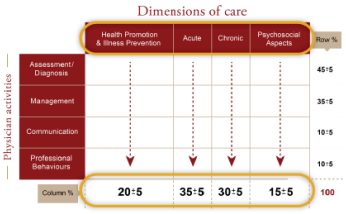


Figure 2: Physician Activities



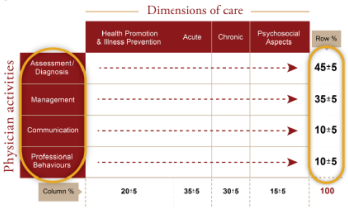
Dimensions of Care



Reflects the focus of care for the patient, family, community and/or population:

- **Health Promotion and Illness Prevention:** The process of enabling people to increase control over their health and its determinants, and thereby improve their health. Illness prevention covers measures not only to prevent the occurrence of illness, such as risk factor reduction, but also to arrest its progress and reduce its consequences once established. This includes, but is not limited to screening, periodic health exam, health maintenance, patient education and advocacy, and community and population health.
- **Acute:** Brief episode of illness within the time span defined by initial presentation through to transition of care. This dimension includes but is not limited to urgent, emergent, and life-threatening conditions, new conditions, and exacerbation of underlying conditions.
- **Chronic:** Illness of long duration that includes but is not limited to illnesses with slow progression.
- **Psychosocial Aspects:** Presentations rooted in the social and psychological determinants of health and how these can impact on wellbeing or illness. The determinants include but are not limited to life challenges, income, culture, and the impact of the patient's social and physical environment.

Physician Activities



Reflects the scope of practice and behaviours of a physician practicing in Canada:

- **Assessment/Diagnosis:** Exploration of illness and disease using clinical judgment to gather, interpret and synthesize relevant information that includes but is not limited to history taking, physical examination and investigation.
- **Management:** Process that includes but is not limited to generating, planning, organizing safe and effective care in collaboration with patients, families, communities, populations, and other professionals (e.g., finding common ground, agreeing on problems and goals of care, time and resource management, roles to arrive at mutual decisions for treatment, working in teams).
- **Communication:** Interactions with patients, families, caregivers, other professionals, communities and populations. Elements include but are not limited to relationship development, intra-professional and inter-professional collaborative care, education, verbal communication (e.g., using the patient-centered interview and active listening), non-verbal and written communication, obtaining informed consent, and disclosure of patient safety incidents.
- **Professional Behaviours:** Attitudes, knowledge, and skills relating to clinical and/or medical administrative competence, communication, ethics, as well as societal and legal duties. The wise application of these behaviours demonstrates a commitment to excellence, respect, integrity, empathy, accountability and altruism within the Canadian health-care system. Professional behaviours also include but are not limited to self-awareness, reflection, life-long learning, leadership, scholarly habits and physician health for sustainable practice.

APPENDIX D:
Internal Structure: New Blueprint

The Medical Council of Canada (MCC) undertook a strategic review of its assessment processes with a clear focus on their purposes and objectives, their structure and alignment with the MCC's major stakeholder requirements. The review addressed current trends in medical education, regulation and assessment. The review also considered the role and purpose of the MCC's examinations in meeting the current and future needs of medical regulatory authorities (MRAs), the public and other stakeholders. In addition to focusing on the reassessment and realignment of the MCC's exams, a key recommendation focused on validating and updating the blueprints for both components of the MCC Qualifying Examination (MCCQE).

As part of its commitment to adhere to best practices in medical education and assessment, the MCC undertook a blueprint project to review and establish an evidence-based approach for identifying the competencies that physicians will be expected to demonstrate and be assessed on at two decision points: (1) entry into residency and (2) entry into independent practice. The purpose is to ensure that critical core competencies, knowledge, skills and behaviours for safe and effective patient care in Canada are being appropriately assessed for the two decision points. The rigorous and consultative process of how the Blueprint was developed can be found *here*.

The new Blueprint offers the MCC the opportunity to assess fundamental core competencies required of physicians practising in Canada at various points along their careers, regardless of specialty, and considers the performance across two broad categories, Dimensions of Care and Physician Activities. The internal structure of the MCCQE Part I can be revealed, to some degree, through the evaluation of the correlations among the Blueprint subscores. Correlating the two categories (and their embedded domains) can help one understand how closely the exam conforms to the construct of interest. Correlations among subscores were examined using the data from 4,166 examinees who took the MCCQE Part I in the April 2018 administration.

Table 1: Correlation matrices among subscores
in the four domains of Dimensions of Care and Total Scores

	Total Score	Health Promotion	Acute	Chronic	Psychosocial Aspects
Total Score	1				
Health Promotion	0.84	1			
Acute	0.91	0.66	1		
Chronic	0.86	0.64	0.68	1	
Psychosocial Aspects	0.67	0.53	0.51	0.48	1

Table 2: Correlation matrices among subscores
in the four domains of Physician Activities and Total Scores

	Total Score	Assessment / Diagnosis	Management	Communication	Professional Behaviours
Total Score	1				
Assessment / Diagnosis	0.91	1			
Management	0.92	0.74	1		
Communication	0.67	0.50	0.55	1	
Professional Behaviours	0.67	0.49	0.55	0.47	1

Table 3: Correlation matrices among subscores
in Physician Activities and in Dimensions of Care

	Health Promotion	Acute	Chronic	Psychosocial Aspects
Assessment / Diagnosis	0.72	0.87	0.81	0.52
Management	0.79	0.84	0.80	0.58
Communication	0.64	0.54	0.53	0.61
Professional Behaviours	0.59	0.55	0.51	0.66

As indicated in Tables 1 to 3, all subscores classified by either Dimensions of Care or Physician Activities were found to be significantly, positively correlated with one another. The highest correlation was found with the Total Score. This suggests that the MCCQE Part I seems to measure an essentially single dominant underlying construct (i.e., basic medical knowledge and clinical skills that the MCCQE Part I is designed to measure). Furthermore, this provides some preliminary evidence to support the assumption of unidimensionality underlying the use of the Rasch model used to assemble and score the exam. It should be noted that the magnitude of correlations may be affected by the number of items in each domain. All correlations presented in Tables 1 to 3 were statistically significant at $p<0.001$.

Tables 4 to 6 present the disattenuated correlations between the domains of the Blueprint. The disattenuated correlation is based on their observed correlation adjusted for reliability of the domains and it indicates what their correlation would be after correction for measurement error. The reliability of a set of measures is the proportion of observed variance not due to measurement error. However, it is important to note that disattenuation (a) is not a substitute for precise measurement, (b) does not change the quality of the measures or their predictive power, (c) is not directly comparable with uncorrected correlations, and (d) is not suitable for statistical hypothesis testing (Muchinsky, 1996).

Table 4: Correlations (adjusting for reliability) among sub-scores in Dimensions of Care

	Reliability	Health Promotion	Acute	Chronic	Psychosocial Aspects
Health Promotion	0.64	1.00[1]			
Acute	0.75	0.95	1.00[1]		
Chronic	0.67	0.97	0.96	1.00[1]	
Psychosocial Aspects	0.46	0.98	0.87	0.87	1.00[1]

[1] Correlations originally greater than 1.0

Table 5: Correlations (corrected for attenuation) among subscores in Physician Activities

	Reliability	Assessment / Diagnosis	Management	Communication	Professional Behaviours
Assessment / Diagnosis	0.75	1.00[1]			
Management	0.74	0.99	1.00[1]		
Communication	0.44	0.86	0.97	1.00[1]	
Professional Behaviours	0.47	0.82	0.93	1.00[1]	1.00[1]

[1] Correlations originally greater than 1.0

Table 6: Correlations (corrected for attenuation) among subscores in Dimensions of Care and Physician Activities

	Health Promotion	Acute	Chronic	Psychosocial Aspects
Assessment /Diagnosis	1.00[1]	1.00[1]	1.00[1]	0.88
Management	1.00[1]	1.00[1]	1.00[1]	0.99
Communication	1.00[1]	0.94	0.98	1.00[1]
Professional Behaviours	1.00[1]	0.93	0.91	1.00[1]

[1] Correlations greater than 1.0

All disattenuated correlations presented in Tables 4 to 6 were statistically significant at $p<0.001$. The greatest impact of adjusting for reliability was in three domains: Psychosocial aspects, Communication and Professional Behaviours. The disattenuated correlations tell us whether the correlation between two sets of measures is low because of measurement error or because the two sets are really uncorrelated (Muchinsky, 1996). The large difference between observed and disattenuated correlations for these three domains suggests that they are indeed correlated; however, measurement error is lowering the correlation below the level it would have reached had the measures been precise. Disattenuated values greater than 1.00 indicate that measurement error is not randomly distributed. In such instances, Muchinsky (1996) suggests reporting these values as 1.00.

APPENDIX E: Generalizability study

Table 1: Average, minimum, and maximum variance across multiple test forms for Language, by effect type and cohort

Design	Cohort	Effect[1]	Variance		
			Average	Min	Max
item x (people:language)	CMG, First time takers	i	0.04	0.03	0.05
		l	0.00	0.00	0.00
		p:l	0.00	0.00	0.00
		il	0.01	0.00	0.01
		ip:l	0.15	0.14	0.16
	First time takers	i	0.04	0.03	0.05
		l	0.00	0.00	0.00
		p:l	0.00	0.00	0.01
		il	0.00	0.00	0.01
		ip:l	0.16	0.15	0.17
	All candidates	i	0.04	0.03	0.05
		l	0.00	0.00	0.00
		p:l	0.01	0.00	0.01
		il	0.00	0.00	0.01
		ip:l	0.17	0.15	0.17

[1] i: item; l: language (English/French); p (people or candidates)

Table 2: Average, minimum, and maximum variance across multiple test forms for Gender by effect type and cohort

Design	Cohort	Effect[1]	Variance		
			Average	Min	Max
item x (people:gender)	CMG, First time takers	i	0.04	0.03	0.05
		g	0.00	0.00	0.00
		p:g	0.00	0.00	0.00
		ig	0.00	0.00	0.00
		ip:g	0.16	0.14	0.16
	First time takers	i	0.04	0.03	0.05
		g	0.00	0.00	0.00
		p:g	0.00	0.00	0.01
		ig	0.00	0.00	0.00
		ip:g	0.16	0.15	0.17
	All candidates	i	0.04	0.04	0.05
		g	0.00	0.00	0.00
		p:g	0.01	0.00	0.01
		ig	0.00	0.00	0.00
		ip:g	0.17	0.16	0.17

[1] i: item; l: gender (Male/Female); p (people or candidates)