Medical Council of Canada Qualifying Examination (MCCQE) Part I

MCCQE PART I ANNUAL TECHNICAL REPORT 2019

APR. 2019 – JAN. 2020

19 12
MEDICAL COUNCIL
OF CANADA
LE CONSEIL MÉDICAL
DU CANADA

Table of Contents

List of Tables and Figures

Preface

This report summarizes the fundamental psychometric characteristics, test development, test publishing, and test administration activities of the Medical Council of Canada Qualifying Examination (MCCQE) Part I. Candidate performance data on the exam in 2019/2020[1] 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 five administrations in 2019/2020 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, members of MCC's Council, 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.

[1] 2019/2020 - includes data from April 2019, July 2019, August 2019, October 2019 and January 2020 exam windows.

The Medical Council of Canada (MCC) undertook a strategic review of its assessment processes with a 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 function 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 Blueprint for both components of the MCC Qualifying Examinations (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, ruling on candidate special cases and reconsiderations, and the approval of results.

2. Exam development

In this section, we describe the exam Blueprint, exam specifications and constraints, 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:

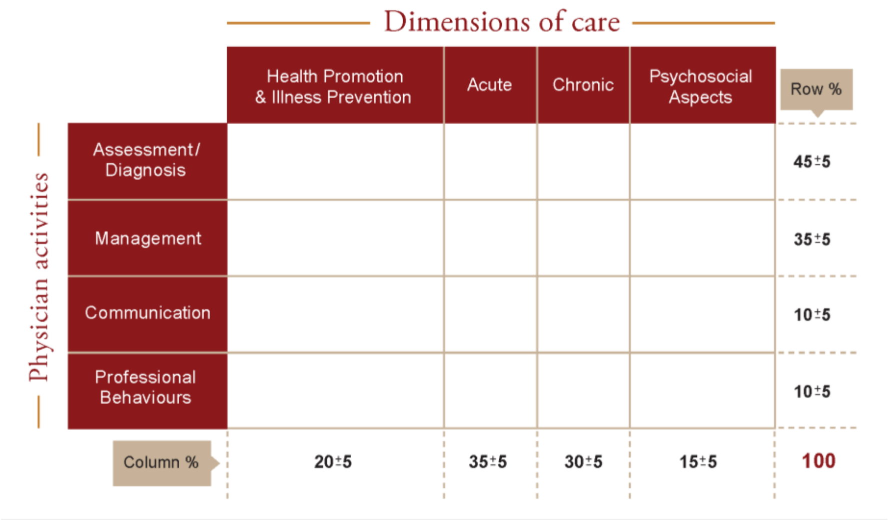
- **Assessment/Diagnosis**: exploration of illness and disease using clinical judgment to gather, interpret and synthesize relevant information that includes but is not limited to history taking, physical examination and investigation.
- **Management**: process that includes but is not limited to generating, 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-centred 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, lifelong 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



	Dimensions of care				
Physician activities	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 and a CDM component– content and psychometric specifications are considered at the total test level.

2.2.1 Content specifications

Table 1 above 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 child-bearing age and the frail elderly
Gender	Male, female	Balance evenly (min. 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 and are not penalized for guessing. 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. If relevant to a question, the candidate will be presented with the normal lab values directly in the question stem.

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 short-answer 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.

Similar to the MCQ section, 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. If relevant to a question, the candidate will be presented with the normal lab values directly in the case or question stem.

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 test form. The target TIF is used to balance multiple test 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.

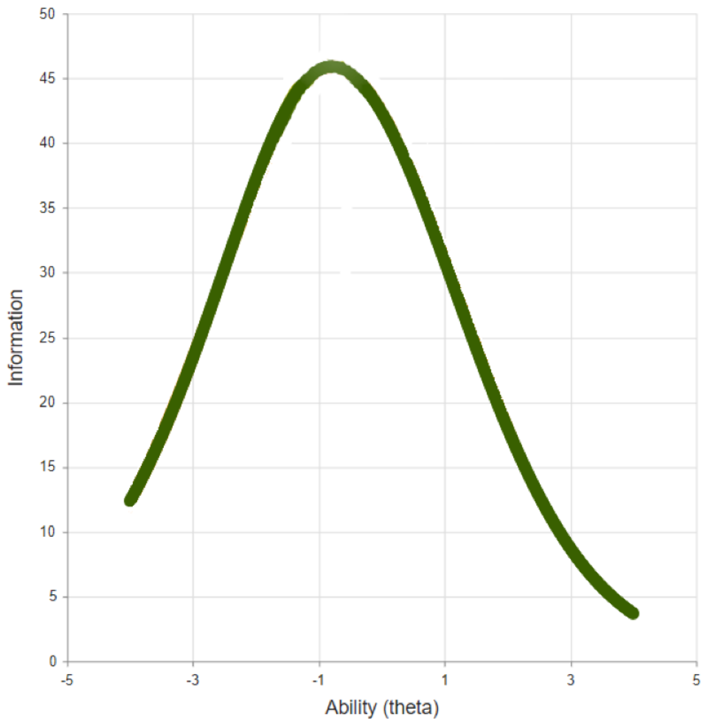


Figure 1. Target test information function

2.3 ITEM DEVELOPMENT

For the MCQ content, six specialty test committees create and approve exam content. For the CDM content, one multidisciplinary test committee develops exam content. In 2019, the MCCQE Part I Test Development Officers (TDO) held a content development workshop that served two purposes; increase the quantity of new CDM content and educate additional item writers in the domain of CDM content. The difference in the CDM Test Committee composition and process is described below in the 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 educates item writers, instructs members on the blueprint and objectives, supports the TDOs in identifying content gap areas, and is a consistent subject matter expert across all test 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 (i.e., subject matter experts) to ensure that the content is defensible. If so, the items are available for use in future administrations of the exam. 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 2019 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 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 the 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, conducts a copy edit, 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 the 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 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.

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 are 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 CDM has been guided by 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 the CDM Test 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). In 2019, these items were shared between adjacent test forms, and they ensured score comparability across test forms.

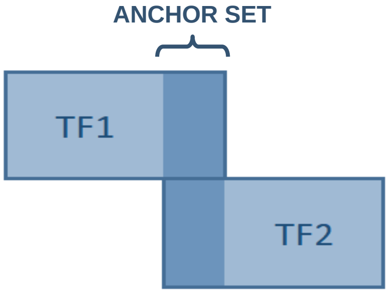


Figure 2. Test form representation

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.

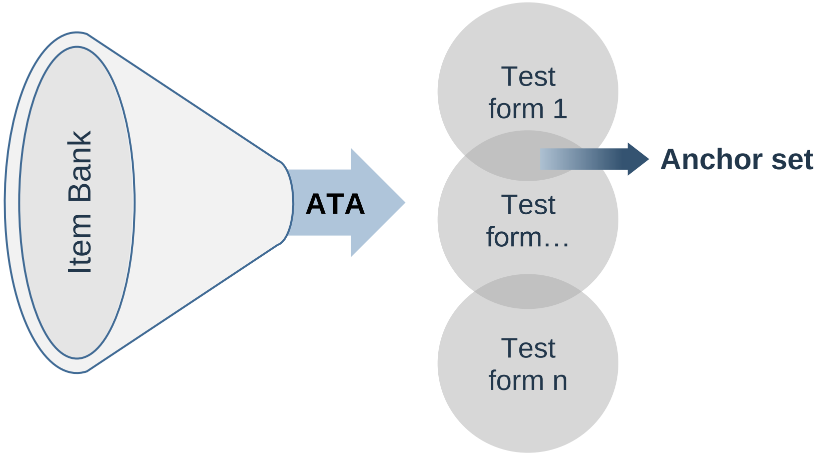


Figure 3: Automated test assembly procedure

The TIF for each of the test forms was inspected. The maximum information for each form was 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 correlates with 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

Starting in 2019, the MCCQE Part I was delivered in Canada and internationally in over 180 countries through a vendor, Prometric. Prometric is an internationally recognized organization with more than 20 years' experience in exam development and administration for professional, high-stakes examination sectors. The change to Prometric ensures broader access for candidates to take the MCCQE Part I.

In 2019/2020, the MCCQE Part I was offered during five test windows in April, July, August, October, and January 2020. The test windows occur over a two-to-four-week period, in Prometric testing centres, which exceed 14,000 locations worldwide. To meet candidate capacity demand in Canada, the April administration uses additional testing centres in university/college computer labs. In both test centre modalities, Prometric staff delivers the exam, follows strict security protocols, and monitors the exam through the Surpass exam delivery system.

The exam may be taken in either English or French at any test centre; however, staff and technical support may be limited to a specific language. In Canada, support in both official languages occurs at the Ottawa, Montreal, and Quebec City test centres.

3.2 EXAM SECURITY

The MCC takes several measures to safeguard exam security. Test publishing processes are well established, test centre guidelines and security protocols are shared and reviewed with each test centre 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 is standardized and delivered by Prometric. Prometric conducts yearly training with all site administrators to communicate enhancements to security 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 an MCC examination have legal and professional responsibilities. The MCC also has a responsibility to candidates and to the Canadian population 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*. 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 test centre administrator may immediately terminate the candidate's exam. The test centre administrator is required to produce a full report (candidate procedure 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 or 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 *free online materials* through the MCC's website. A *self-education program* for physicians to learn about communication and cultural competencies required in Canada is offered, through online modules, on the physiciansapply.ca website.

Candidates may also test their knowledge by purchasing a full-length Preparatory Examination or Multiple-Choice Questions and Clinical Decision-Making practice tests. More information on the *MCC's preparatory products* can be found on the MCC's website.

3.4 QUALITY ASSURANCE

After each exam day administration, MCC's database is updated with each candidate's response file. An initial system validation is done to ensure an accurate and complete candidate response file is received.

A second validation is completed at the end of the administration where 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 the 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, is informed about administrative issues, rules on special candidate cases, and approves exam results.

The MCC releases candidates' 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 A), the official results document, and the Supplemental Information Report (Appendix B) 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. **Generalization**: 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 (Generalization), 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).

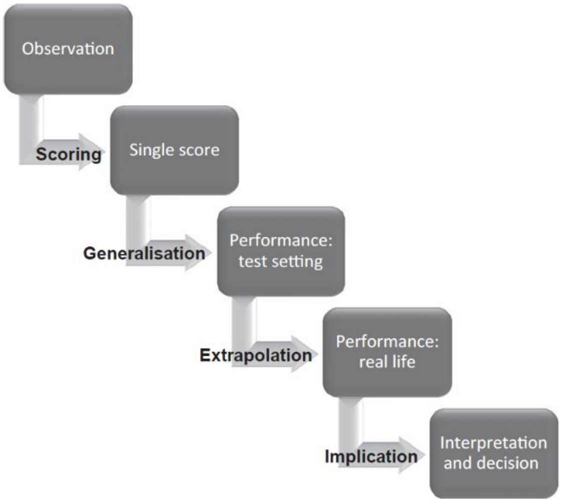


Figure 4: Key elements in Kane's argument-based approach to validation: Inferences from observation to decision (Source: Cook, 2015, page 564)

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.

Sources of Evidence	Datum	Warrant	Backing
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 unintended potential 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-irrelevant variance.
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	Write-in 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 international medical graduates [IMGs]). Detailed information can be found in section 6.3 of this report.

Sources of Evidence	Datum	Warrant	Backing
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 an 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 content	Item parameters	Test forms are comparable in levels 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 between 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 C). 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.

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	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

Sources of Evidence	Datum	Warrant	Backing
		evidence of convergent validity	into a Canadian residency program. A significant correlation (r=.55, p<.0001) was obtained between scores on the MCCQE Part I and the NAC Examination 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 panellists, 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 panellists; careful training of panellists, standard-setting methodology followed best practice (Bookmark and Hofstee methods); and feedback of the panellists post standard-setting exercise. Internal evidence included the consistency of the panellists 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 were repeated by many different panels of similar composition. More information on the standard-setting procedure can be found *here*.

5. Psychometric analyses

In 2019/2020, the MCCQE Part I was offered during five test windows in April, July, August, October and January 2020.

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 remaining sessions.

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 Canadian medical graduates taking the exam for the first time. 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-measure 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 partial credit model (Masters, 1982), a model of the family of Rasch models, 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, the partial credit model is used to establish the difficulty level for the transitions (i.e., steps) between successive item scores. These transitions are modelled using step parameters (or step thresholds), and are expected 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 2 on an item requires higher overall ability than a score of 1. 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 the correct answer
- Item mean square outfit < 0.5
- Item mean square outfit > 2.0.
- Low category score frequency N <10
- Disordered Threshold (CDM only)
- Average ability not increasing (CDM 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. The fall administrations are processed using the same item difficulty estimates from spring so that scores are on the same scale and thus comparable.

5.2 ITEM CALIBRATION

Previous research studies (De Champlain, Boulais, & Dallas, 2016; 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 IRT 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 for dichotomous data, the probability of a correct response on an item is modelled 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, 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 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 April 2019 MCCQE Part I, items were recalibrated maintaining the scale established in 2018. Data from Canadian Medical Graduates (CMGs) first-time test takers was used for this process. First, the parameters for all the 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 SMEs review all flagged items (in Step 1) and provide decision on which items to remove from scoring and calibration, items are recalibrated excluding those items. A final set of 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 SMEs review all flagged items (in Step 1) and provide 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

After items are calibrated and/or vetted by SMEs, items parameters are used for estimating candidate ability of all candidates. Item parameters are fixed in the estimation process and only the level of candidate ability is estimated.

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 partial credit model (Masters, 1982) to score candidates' exam responses. While raw score data (scores of the 1/0 types) 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 predefined scoring keys. To ensure correctness in the scoring process, a rigorous QA process is implemented at this stage: test items are independently scored (using the predefined scoring keys) by two statistical analysts, using two different statistical software. Results are compared and after 100% match, they are reviewed by the psychometrician to ensure reasonableness.

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 predetermined 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. The process of attributing scores to the write-in items is similar to the MCQ and CDM short-menu items described above. In other words, it goes through the same rigorous process of QA.

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, successive integer scores are assigned, also called category scores. For example, a candidate selecting two out of three correct answers would receive two points.

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 -3.0 and +3.0 although values beyond the latter can occur. A candidate who obtains a score of -3.0 would demonstrate very little ability in regard to the specialty areas being assessed whereas a candidate who obtains a score of +3.0 would demonstrate strong ability.

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 exercise[2] 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 panellists 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.

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

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 logit scale. Though the logit 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 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;

And $X_i =$ 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 ability score, estimated using the partial credit model, 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 A and B, 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 five administrations in 2019/2020 is summarized in this section. When applicable, historical data from previous years are included for reference.

6.1 CANDIDATE COHORTS

The 2019/2020 MCCQE Part I includes data from April 2019, July 2019, August 2019, October 2019 and January 2020 windows. The exam was administered in a four-and-a-half-week window (April 12 to May 15), a two-and-a-half-week window (July 8 to July 25), a four-week window (August 27 to September 23), a two-and-a-half-week window (October 28 to November 15) and 4-week window (Jan 20 – Feb 9, 2020). A total of 7,907 candidates challenged the exam across 57 countries. Of the total number of candidates who took the examination in 2019/2020, 29 candidates received a No Standing. 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 (Percentages do not total 100 due to rounding).

	Apr. 2019		July 2019		Aug. 2019		Oct. 2019		Jan. 2020	
Group	N	%	N	%	N	%	N	%	N	%
CMG first-time test takers	2,821	68.3	9	0.9	17	1.6	6	0.7	17	2.3
CMG repeat test takers	45	1.1	16	1.6	29	2.7	37	4	12	1.6
IMG first-time test takers	1,035	25.0	831	81.7	849	78.8	676	73.2	554	73.3
IMG repeat test takers	232	5.6	161	15.8	182	16.9	205	22.2	173	22.9
Total	**4,133**	**100**	**1,017**	**100**	**1,077**	**100**	**924**	**100**	**756**	**100**

6.2 OVERALL EXAM RESULTS

Table 8 summarizes pass rates for the 2019/2020 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 29 candidates who received a No Standing.

Table 8: Exam results – 2019/2020

		Exam Results				
		April 2019	July 2019	Aug. 2019	Oct. 2019	Jan. 2020
CMG first-time test takers	N	2,812	9	17	6	17
	M	263	256	269	246	265
	SD	21.4	23.1	23.1	16.0	19.7
	Min.	190	213	234	226	240
	Max.	340	285	319	268	303
	Pass Rate (%)	**96.6**	**88.9**	**100**	**100**	**100**
CMG repeat test takers	N	45	16	29	36	12
	M	239	243	235	238	223
	SD	17.1	21.5	18.0	16.2	18.9
	Min.	208	210	204	209	199
	Max.	275	287	277	279	254
	Pass Rate (%)	**75.6**	**81.3**	**72.4**	**75**	**50**
IMG first-time test takers	N	1,025	830	847	676	551
	M	228	226	226	221	221
	SD	35.3	33.5	32.0	36.4	34.9
	Min.	109	106	108	100	102
	Max.	321	320	323	302	305
	Pass Rate (%)	**56.3**	**55.3**	**56.4**	**47.2**	**48.1**
IMG repeat test takers	N	232	159	181	205	173
	M	223	224	216	218	219
	SD	19.5	19.6	20.5	22.9	21.8
	Min.	163	177	137	142	132
	Max.	274	268	263	227	259
	Pass Rate (%)	**48.3**	**49.7**	**34.3**	**37.6**	**42.2**
All candidates	N	4,114	1,014	1,074	923	753
	M	252	226	226	221	222
	SD	30.5	31.7	30.7	33.4	32.5
	Min.	109	106	108	100	102
	Max.	340	320	323	302	305
	Pass Rate (%)	**83.6**	**55.1**	**53.8**	**46.5**	**47.9**

Figure 5 displays the total score distribution on the reported score scale for all candidates in the five sessions and total. Overall, the total score performance of the April cohort was better than the other four cohorts.

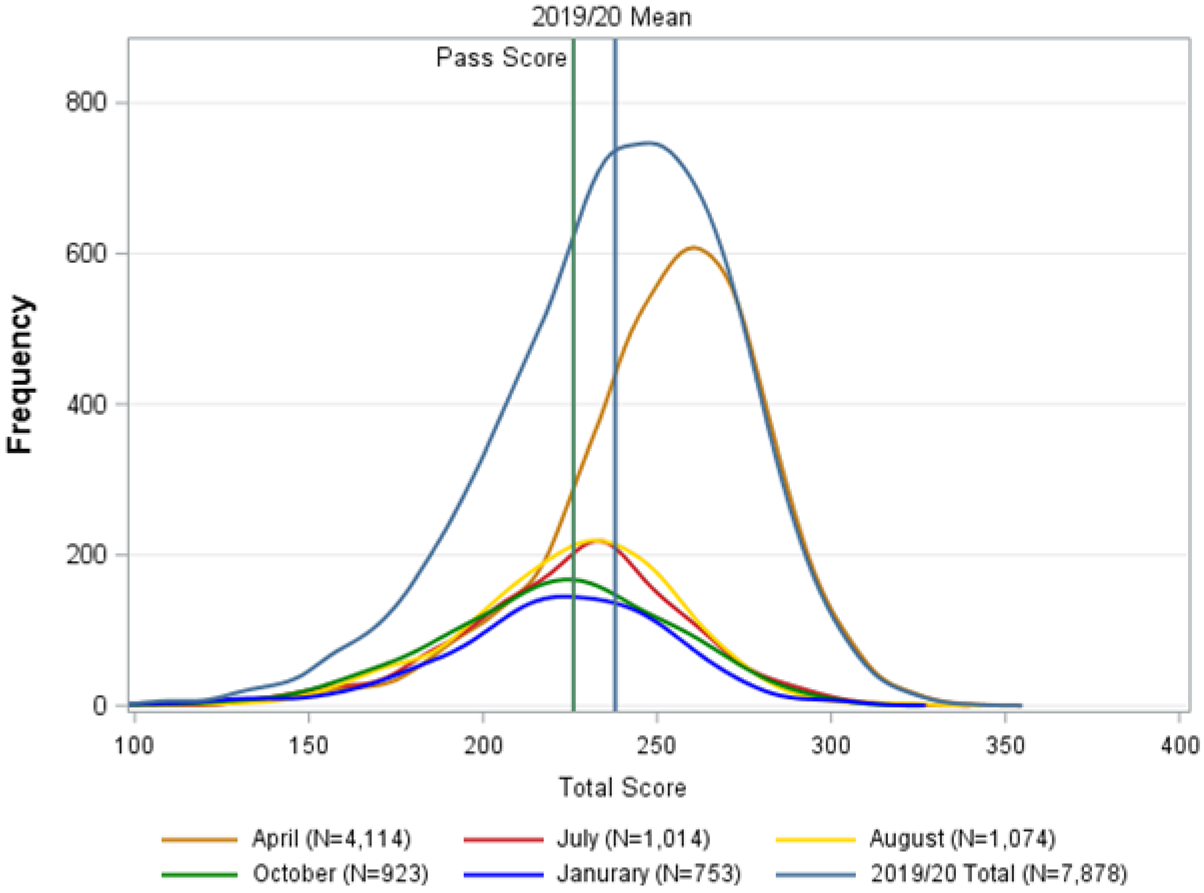


Figure 5: Total exam score distributions – 2019/2020

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) of ability estimates 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 estimation of SE varies with ability levels in IRT whereas the most common estimation methods of the classical SEM do not.

Figures 6 through 10 display scatter plots of SE values along the reported score scale for the five 2019/2020 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.

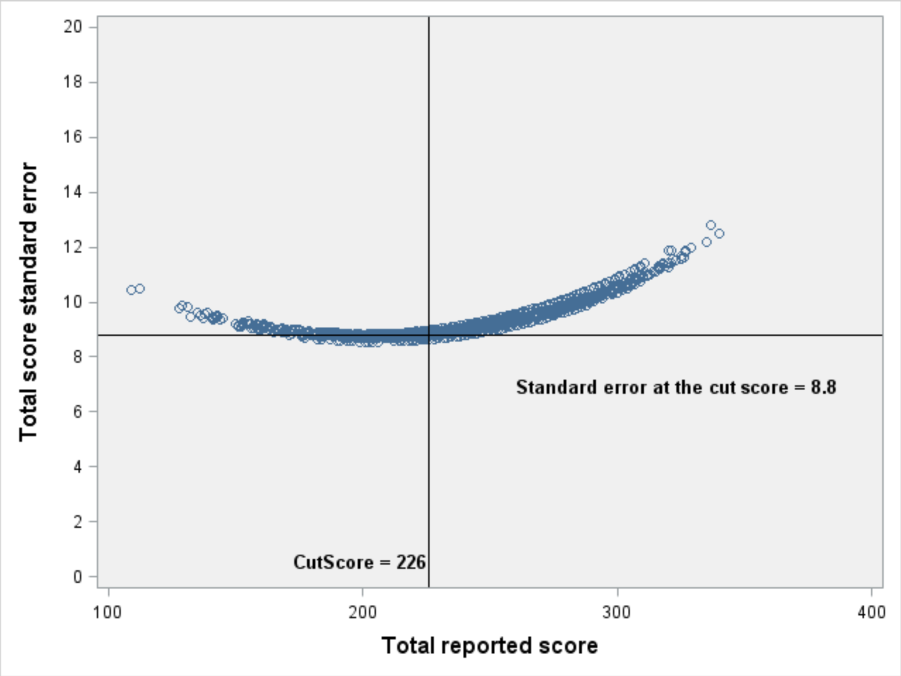


Figure 6: Standard error of ability estimates – April 2019

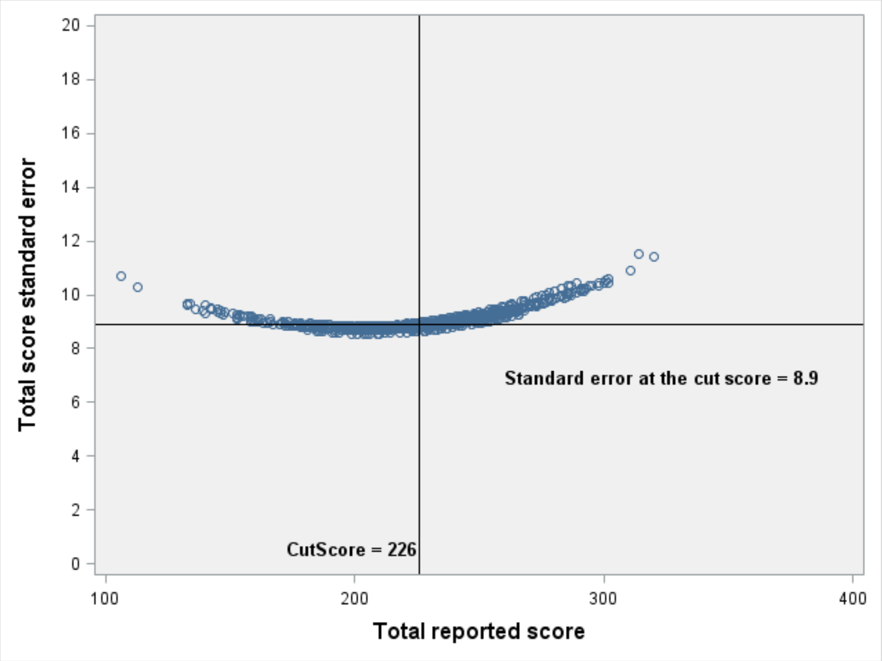


Figure 7: Standard error of ability estimates – July 2019

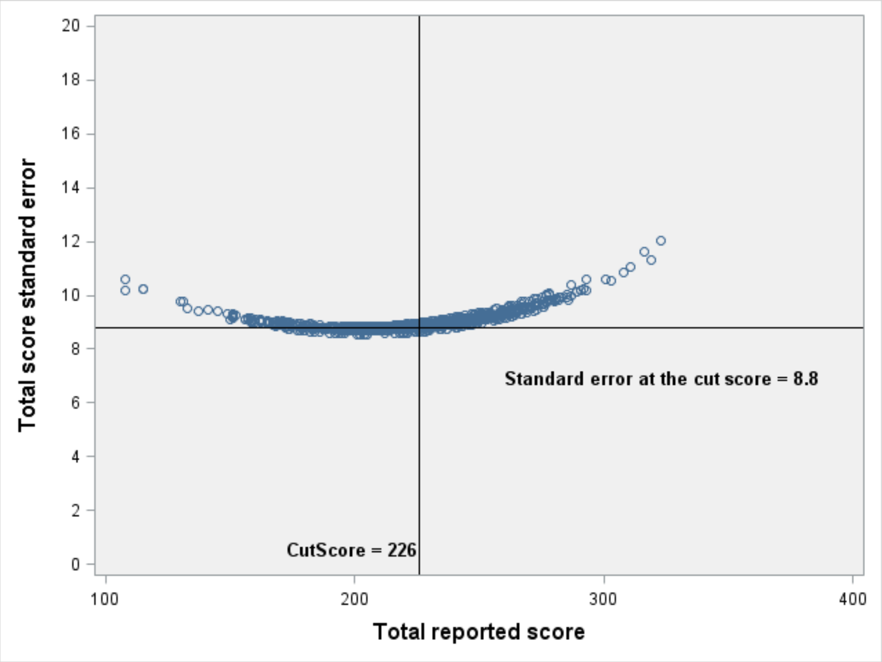


Figure 8: Standard error of ability estimates – August 2019

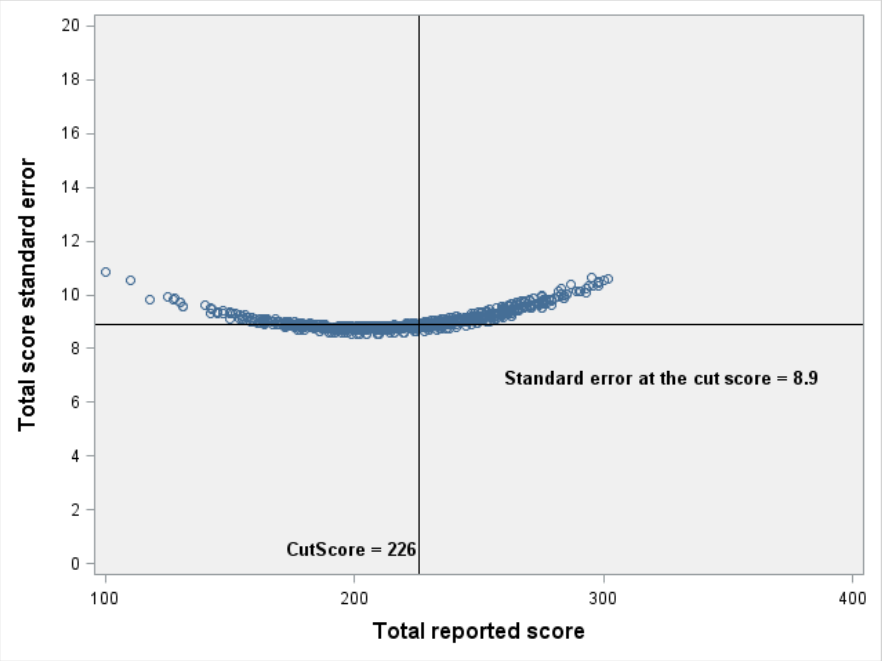


Figure 9: Standard error of ability estimates – October 2019

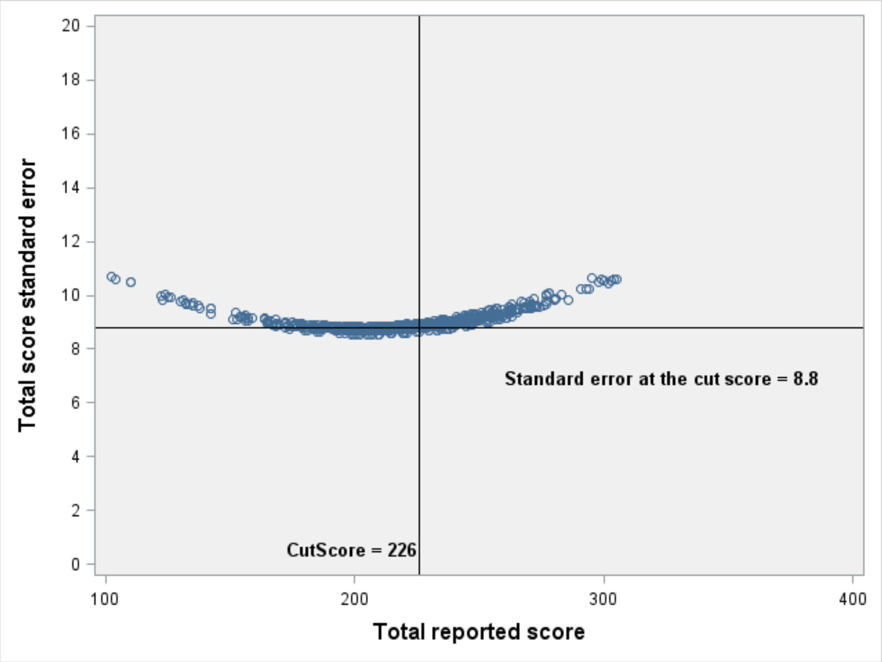


Figure 10: Standard error of ability estimates – January 2020

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 five administrations of 2019/2020 indicate reliable and valid pass/fail decisions based on MCCQE Part I scores. Table 9 is based on data from 4114 candidates in the April session, 1,014 in the July session, 1,074 in the August session, 923 in the October session, and 753 in the January session.

Table 9: Reliability estimates, standard errors of measurement, decision consistency and decision accuracy indices for each administration of 2019/2020

	April	July	August	October	January
Reliability estimate[1]	0.90	0.92	0.91	0.92	0.91
Average SEM (total score)	9.4	9.0	9.0	9.0	9.0
Decision consistency	0.92	0.87	0.87	0.87	0.86
False positive	0.03	0.07	0.07	0.06	0.07
False negative	0.04	0.06	0.06	0.06	0.07
Decision accuracy	0.95	0.91	0.91	0.91	0.90
False positive	0.02	0.05	0.05	0.05	0.05
False negative	0.03	0.04	0.05	0.05	0.05

[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 2019/2020 administrations. The range of domain subscores is presented graphically in Figures 11 to 15. The graphs show the domain subscore for each of the eight domains. The boxes for each domain indicate the range of scores for 50 per cent of the candidates’ domain subscores. The vertical line represents the median or 50th percentile subscore. The remaining 50 per cent of domain subscores are shown to the right or the left of the box as a line (25 per cent to the right and 25 per cent to the left).

The legend for each of the subscores displayed on the figures follows:

Dimensions of Care:	***Physician Activities:***
HEALTHP = Health Promotion & Illness Prevention	PSYCHS = Psychosocial Aspects
ACUTE = Acute	MGMT = Management
CHRONIC = Chronic	COMM = Communication
PSYCHS = Psychosocial Aspects	PROFB = Professional Behaviours

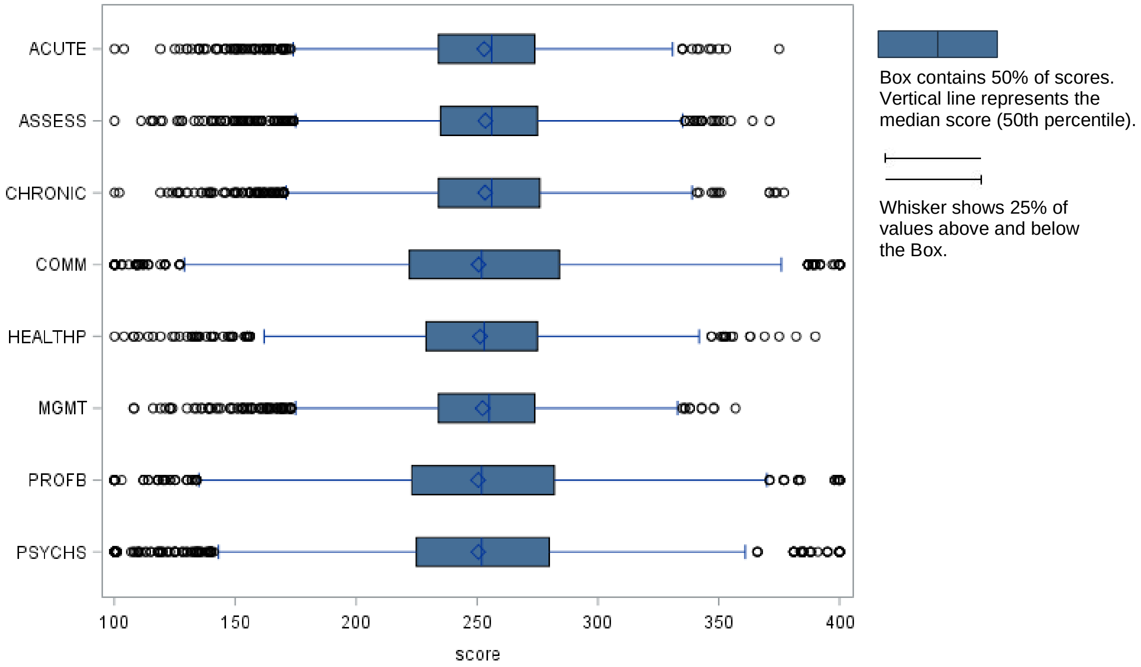


Figure 11: Domain subscore for the April 2019

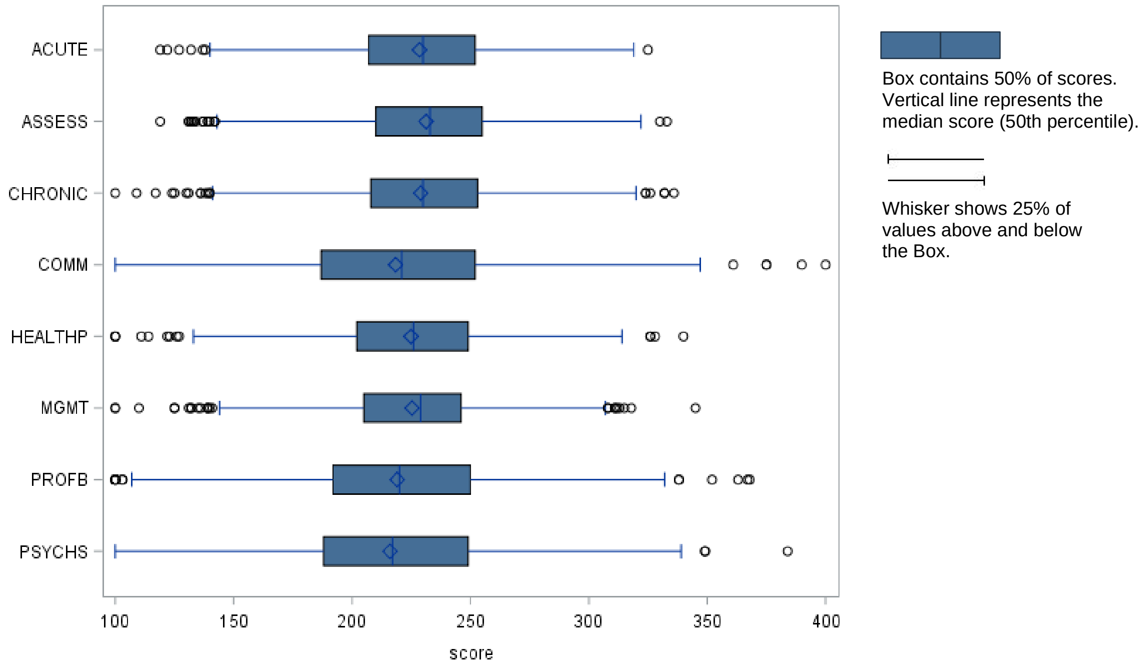


Figure 12: Domain subscore for the July 2019

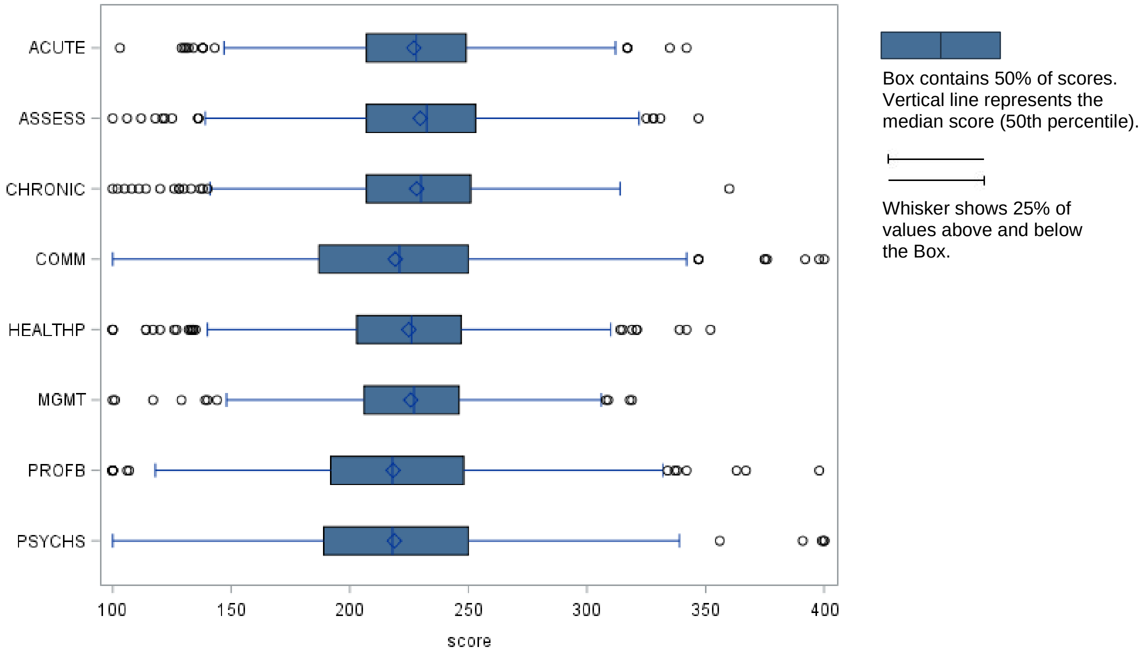


Figure 13: Domain subscore for the August 2019

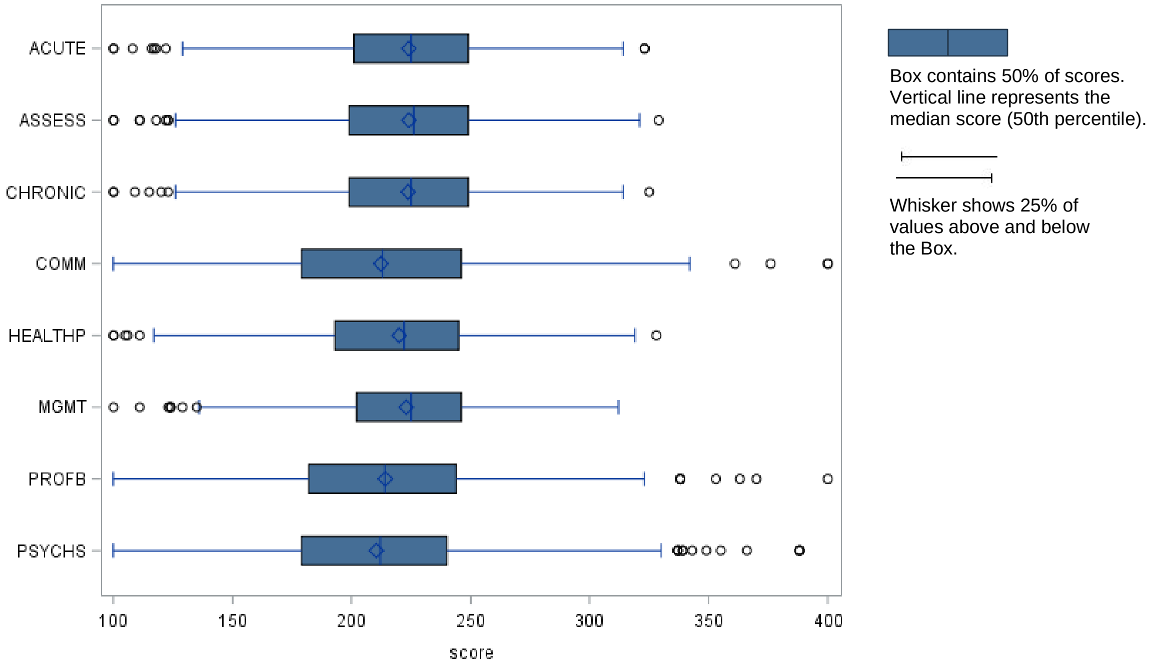


Figure 14: Domain subscore for the October 2019

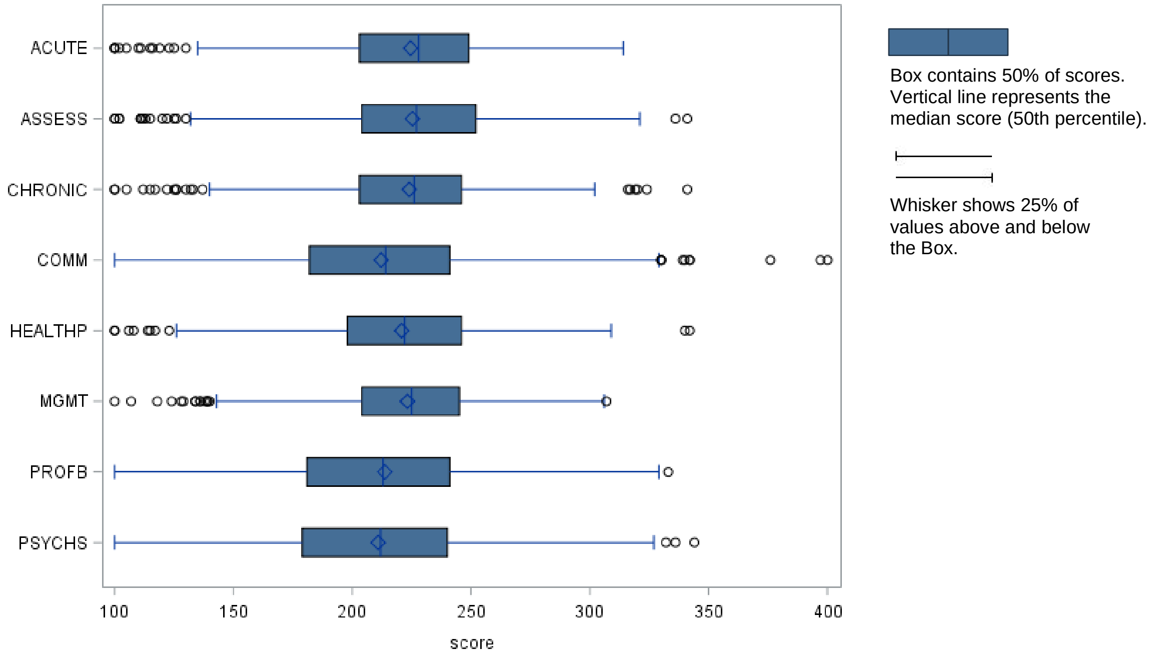


Figure 15: Domain subscore for the January 2020

6.6 HISTORICAL PASS RATES

Historical pass rates are presented in this section. Table 10 shows the pass rates for 2017 to 2019/2020 by group.

Table 10: Spring 2017 to January 2020 pass rates

	2017		2018		2019/2020	
	N	Pass rate	N	Pass rate	N	Pass rate
CMG first-time test takers	2802	95	2823	95	2,861	97
CMG repeat takers	156	63	178	67	138	73
IMG first-time test takers	1677	62	1413	62	3,929	53
IMG repeat takers	1264	29	991	24	950	42
TOTAL	**5899**	**71**	**5405**	**73**	**7,878**	**68**

6.7 CANDIDATE SURVEY

For quality improvement purposes, a survey is administered with each exam to gather candidate feedback regarding their test-taking experience. Tables 11 and 12 present the results of the survey questions, post-MCQ session and post-CDM session respectively, to which candidates responded. In Tables 11 and 12, the percentage of Missing is equal to the number of missing answers divided by the sum of the number of missing answers and valid answers.

Table 11: Candidate survey results –
percentages for ratings: Post-MCQ survey

The MCQ section of the MCCQE Part I provided an opportunity for me to demonstrate my level of medical knowledge prior to entering supervised practice.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	198 (5%)	2044 (52%)	1088 (27%)	538 (14%)	93 (2%)	3961 (100%)	170 (4%)
July 2019	121 (13%)	505 (53%)	209 (22%)	97 (10%)	21 (2%)	953 (100%)	64 (6%)
Aug. 2019	152 (15%)	481 (47%)	236 (23%)	109 (11%)	36 (4%)	1014 (100%)	62 (6%)
Oct. 2019	116 (13%)	461 (53%)	194 (23%)	71 (8%)	20 (2%)	862 (100%)	63 (7%)
Jan. 2020	96 (14%)	363 (52%)	163 (23%)	60 (9%)	21 (3%)	703 (100%)	52 (7%)

The test security video was easy to understand and made me aware of my obligations regarding test security and confidentiality.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	1104 (28%)	2246 (57%)	488 (12%)	68 (2%)	36 (1%)	3942 (100%)	189 (5%)
July 2019	357 (38%)	493 (52%)	80 (8%)	12 (1%)	5 (1%)	947 (100%)	70 (7%)
Aug. 2019	403 (40%)	488 (48%)	97 (10%)	13 (1%)	7 (1%)	1008 (100%)	68 (6%)
Oct. 2019	305 (36%)	474 (55%)	69 (8%)	7 (1%)	4 (0%)	859 (100%)	66 (7%)
Jan. 2020	247 (35%)	379 (54%)	62 (9%)	8 (1%)	5 (1%)	701 (100%)	54 (7%)

The exam tutorial was clear, and the review of the functionality was helpful.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	1247 (32%)	2356 (60%)	270 (7%)	48 (1%)	20 (1%)	3941 (100%)	190 (5%)
July 2019	357 (38%)	509 (54%)	71 (8%)	9 (1%)	3 (0%)	949 (100%)	68 (7%)
Aug. 2019	382 (38%)	530 (53%)	74 (7%)	10 (1%)	9 (1%)	1005 (100%)	71 (7%)
Oct. 2019	290 (34%)	511 (59%)	55 (6%)	4 (0%)	1 (0%)	861 (100%)	64 (7%)
Jan. 2020	242 (35%)	403 (58%)	38 (5%)	12 (2%)	4 (1%)	699 (100%)	56 (7%)

The MCQ instructions were clear and easy to follow.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	1124 (29%)	2455 (62%)	272 (7%)	66 (2%)	23 (1%)	3940 (100%)	191 (5%)
July 2019	303 (32%)	542 (57%)	77 (8%)	19 (2%)	6 (1%)	947 (100%)	70 (7%)
Aug. 2019	358 (36%)	531 (53%)	80 (8%)	26 (3%)	9 (1%)	1004 (100%)	72 (7%)
Oct. 2019	255 (30%)	517 (60%)	70 (8%)	15 (2%)	2 (0%)	859 (100%)	66 (7%)
Jan. 2020	207 (30%)	419 (60%)	47 (7%)	22 (3%)	3 (0%)	698 (100%)	57 (8%)

The MCQ content was of high quality (e.g., questions were clearly written and free of any typos or grammatical errors).

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	480 (12%)	2052 (52%)	841 (21%)	482 (12%)	75 (2%)	3930 (100%)	201 (5%)
July 2019	213 (22%)	505 (53%)	147 (16%)	66 (7%)	17 (2%)	948 (100%)	69 (7%)
Aug. 2019	231 (23%)	490 (49%)	171 (17%)	83 (8%)	22 (2%)	997 (100%)	79 (7%)
Oct. 2019	175 (20%)	477 (56%)	134 (16%)	57 (7%)	14 (2%)	857 (100%)	68 (7%)
Jan. 2020	157 (23%)	350 (50%)	135 (19%)	43 (6%)	12 (2%)	697 (100%)	58 (8%)

The images were of high quality (e.g., image was clear and easy to see the details required).

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	627 (16%)	2065 (53%)	566 (14%)	575 (15%)	94 (2%)	3927 (100%)	204 (5%)
July 2019	213 (22%)	478 (50%)	104 (11%)	122 (13%)	31 (3%)	948 (100%)	69 (7%)
Aug. 2019	230 (23%)	493 (49%)	134 (13%)	118 (12%)	25 (3%)	1000 (100%)	76 (7%)
Oct. 2019	185 (22%)	477 (56%)	108 (13%)	73 (9%)	13 (2%)	856 (100%)	69 (7%)
Jan. 2020	177 (25%)	344 (49%)	97 (14%)	65 (9%)	14 (2%)	697 (100%)	58 (8%)

Overall, how would you rate the usability of the examination (including such factors as the screen layout and on-screen functionality)?

	Excellent	Very good	Good	Fair	Poor	Total valid	Missing
April 2019	723 (18%)	1658 (42%)	1157 (29%)	315 (8%)	70 (2%)	3923 (100%)	208 (5%)
July 2019	193 (20%)	384 (41%)	293 (31%)	59 (6%)	15 (2%)	944 (100%)	73 (7%)
Aug. 2019	234 (23%)	401 (40%)	278 (28%)	70 (7%)	14 (1%)	997 (100%)	79 (7%)
Oct. 2019	181 (21%)	350 (41%)	246 (29%)	68 (8%)	9 (1%)	854 (100%)	71 (8%)
Jan. 2020	155 (22%)	269 (39%)	207 (30%)	51 (7%)	11 (2%)	693 (100%)	62 (8%)

How would you rate the performance of the computer and testing application during your exam?

	Excellent	Very good	Good	Fair	Poor	Total valid	Missing
April 2019	761 (19%)	1515 (39%)	1050 (27%)	420 (11%)	176 (4%)	3922(100%)	209 (5%)
July 2019	241 (25%)	380 (40%)	249 (26%)	61 (6%)	16% (2%)	947 (100%)	70 (7%)
Aug. 2019	273 (28%)	405 (41%)	228 (23%)	64 (6%)	22 (2%)	992 (100%)	84 (8%)
Oct. 2019	236 (28%)	343 (40%)	223 (26%)	47 (5%)	7 (1%)	856 (100%)	69 (7%)
Jan. 2020	175 (25%)	271 (39%)	192 (28%)	41 (6%)	13 (2%)	692 (100%)	63 (8%)

Table 12: Candidate survey results –
2019/2020 percentages for ratings: Post-CDM survey

The CDM section of the MCCQE Part I provided an opportunity for me to demonstrate my level of medical knowledge prior to entering supervised practice.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	298 (8%)	2017 (51%)	1051 (27%)	454 (12%)	120 (3%)	3940 (100%)	191 (5%)
July 2019	154 (16%)	508 (53%)	223 (23%)	59 (6%)	23 (2%)	967 (100%)	49 (5%)
Aug. 2019	176 (17%)	498 (49%)	233 (23%)	83 (8%)	22 (2%)	1012 (100%)	64 (6%)
Oct. 2019	138 (16%)	488 (56%)	165 (19%)	66 (8%)	15 (2%)	872 (100%)	53 (6%)
Jan. 2020	118 (17%)	356 (50%)	168 (24%)	54 (8%)	18 (3%)	714 (100%)	41 (5%)

The CDM instructions were clear and easy to follow.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	597 (15%)	2562 (65%)	510 (13%)	224 (6%)	39 (1%)	3932 (100%)	199 (5%)
July 2019	185 (19%)	574 (59%)	145 (15%)	51 (5%)	11 (1%)	966 (100%)	50 (5%)
Aug. 2019	219 (22%)	579 (57 %)	140 (14%)	59 (6%)	14 (1%)	1011 (100%)	65 (6%)
Oct. 2019	164 (19%)	540 (62%)	118 (14%)	41 (5%)	9 (1%)	872 (100%)	53 (6%)
Jan. 2020	157 (22%)	418 (59%)	97 (14%)	34 (5%)	7 (1%)	713 (100%)	42 (6%)

The CDM content was of high quality (e.g., questions were clearly written and free of any typos or grammatical errors).

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	427 (11%)	2138 (55%)	796 (20%)	486 (12%)	69 (2%)	3916 (100%)	215 (5%)
July 2019	166 (17%)	540 (56%)	178 (19%)	71 (7%)	4 (0%)	959 (100%)	57 (6%)
Aug. 2019	206 (20%)	532 (53%)	195 (19%)	68 (7%)	8 (1%)	1009 (100%)	67 (6%)
Oct. 2019	160 (18%)	507 (58%)	138 (16%)	55 (6%)	10 (1%)	870 (100%)	55 (6%)
Jan. 2020	149 (21%)	400 (56%)	110 (15%)	39 (5%)	12 (2%)	710 (100%)	45 (6%)

The images were of high quality (e.g., image was clear and easy to see the details required).

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	Total valid	Missing
April 2019	529 (14%)	2219 (57%)	703 (18%)	385 (10%)	72 (2%)	3908 (100%)	223 (5%)
July 2019	179 (19%)	510 (53%)	162 (17%)	90 (9%)	16 (2%)	957 (100%)	59 (6%)
Aug. 2019	202 (20%)	522 (52%)	185 (18%)	77 (8%)	21 (2%)	1007 (100%)	69 (6%)
Oct. 2019	161 (19%)	486 (56%)	151 (17%)	61 (7%)	9 (1%)	868 (100%)	57 (6%)
Jan. 2020	138 (19%)	401 (56%)	103 (14%)	61 (9%)	9 (1%)	712 (100%)	43 (6%)

Overall, how would you rate the usability of the examination (including such factors as the screen layout and on-screen functionality)?

	Excellent	Very good	Good	Fair	Poor	Total valid	Missing
April 2019	600 (15%)	1473 (38%)	1126 (29%)	509 (13%)	196 (5%)	3904 (100%)	227 (6%)
July 2019	201 (21%)	379 (40%)	283 (30%)	81 (8%)	15 (2%)	959 (100%)	57 (6%)
Aug. 2019	223 (22%)	414 (41%)	254 (25%)	92 (9%)	23 (2%)	1006 (100%)	70 (7%)
Oct. 2019	185 (21%)	356 (41%)	257 (30%)	58 (7%)	12 (1%)	868 (100%)	57 (6%)
Jan. 2020	167 (24%)	273 (38%)	197 (28%)	57 (8%)	16 (2%)	710 (100%)	45 (6%)

How would you rate the site staff availability/helpfulness?

	Excellent	Very good	Good	Fair	Poor	Total valid	Missing
April 2019	1350 (35%)	1604 (41%)	770 (20%)	157 (4%)	31 (1%)	3912 (100%)	219 (5%)
July 2019	412 (43%)	360 (38%)	153 (16%)	29 (3%)	4 (0%)	958 (100%)	58 (6%)
Aug. 2019	442 (44%)	370 (37%)	149 (15%)	37 (4%)	11 (1%)	1009 (100%)	67 (6%)
Oct. 2019	318 (36%)	346 (40%)	159 (18%)	41 (5%)	8 (1%)	872 (100%)	53 (6%)
Jan. 2020	276 (39%)	257 (36%)	139 (20%)	32 (4%)	9 (1%)	713 (100%)	42 (6%)

How would you rate the examination room/computer lab (e.g., physical layout conducive to a high-stakes examination)?

	Excellent	Very good	Good	Fair	Poor	Total valid	Missing
April 2019	864 (22%)	1528 (39%)	1059 (27%)	326 (8%)	121 (3%)	3898 (100%)	223 (6%)
July 2019	312 (33%)	351 (37%)	223 (23%)	45 (5%)	23 (2%)	954 (100%)	62 (6%)
Aug. 2019	304 (30%)	393 (39%)	218 (22%)	70 (7%)	17 (2%)	1002 (100%)	74 (7%)
Oct. 2019	244 (28%)	349 (40%)	210 (24%)	55 (6%)	6 (1%)	864 (100%)	61 (7%)
Jan. 2020	215 (30%)	264 (37%)	174 (25%)	48 (7%)	8 (1%)	709 (100%)	46 (6%)

How would you rate your overall examination experience?

	Excellent	Very good	Good	Fair	Poor	Total valid	Missing
April 2019	302 (8%)	1215 (31%)	1651 (42%)	640 (16%)	100 (3%)	3908 (100%)	223 (5%)
July 2019	123 (13%)	319 (33%)	362 (38 %)	129 (13%)	23 (2%)	956 (100%)	60 (6%)
Aug. 2019	157 (16%)	331 (33%)	372 (37%)	124 (12%)	23 (2%)	1007 (100%)	69 (6%)
Oct. 2019	128 (15%)	292 (34%)	319 (37%)	112 (13%)	17 (2%)	868 (100%)	57 (6%)
Jan. 2020	89 (13%)	243 (34%)	257 (36%)	99 (14%)	19 (3%)	707 (100%)	48 (6%)

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, L. (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.org/10.1111/medu.12678

De Champlain, A., Boulais, A.-P., & Dallas, A. (2016). Calibrating the Medical Council of Canada’s Qualifying Examination Part I using an integrated item response theory framework: a comparison of models and designs. *J Educ Eval Health Prof*, *13*:6. doi.org/10.3352/jeehp.2016.13.6.

Frank, J.R., Snell, L., & Sherbino, J. (2015). *CanMEDS 2015 Physician Competency Framework*. Ottawa: Royal College of Physicians and Surgeons of Canada. Retrieved from: researchgate.net/publication/289254803_CanMEDS_2015_Physician_Competency_Framework

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 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*.

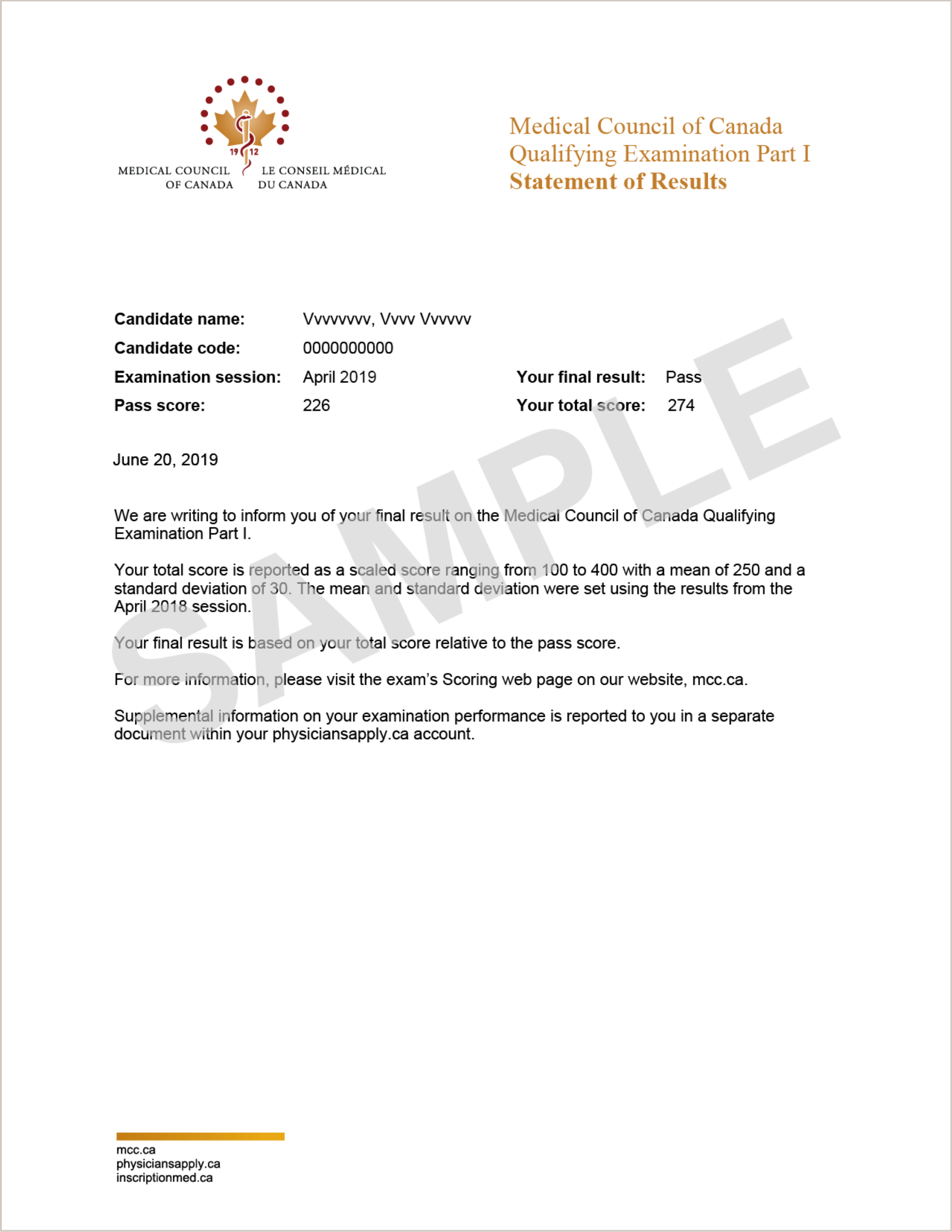
Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174. *dx.doi.org/ 10.1007/BF02296272*.

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 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 2019	**Your final result:**	Pass
Pass score:	226	**Your total score:**	274

June 20, 2019

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.

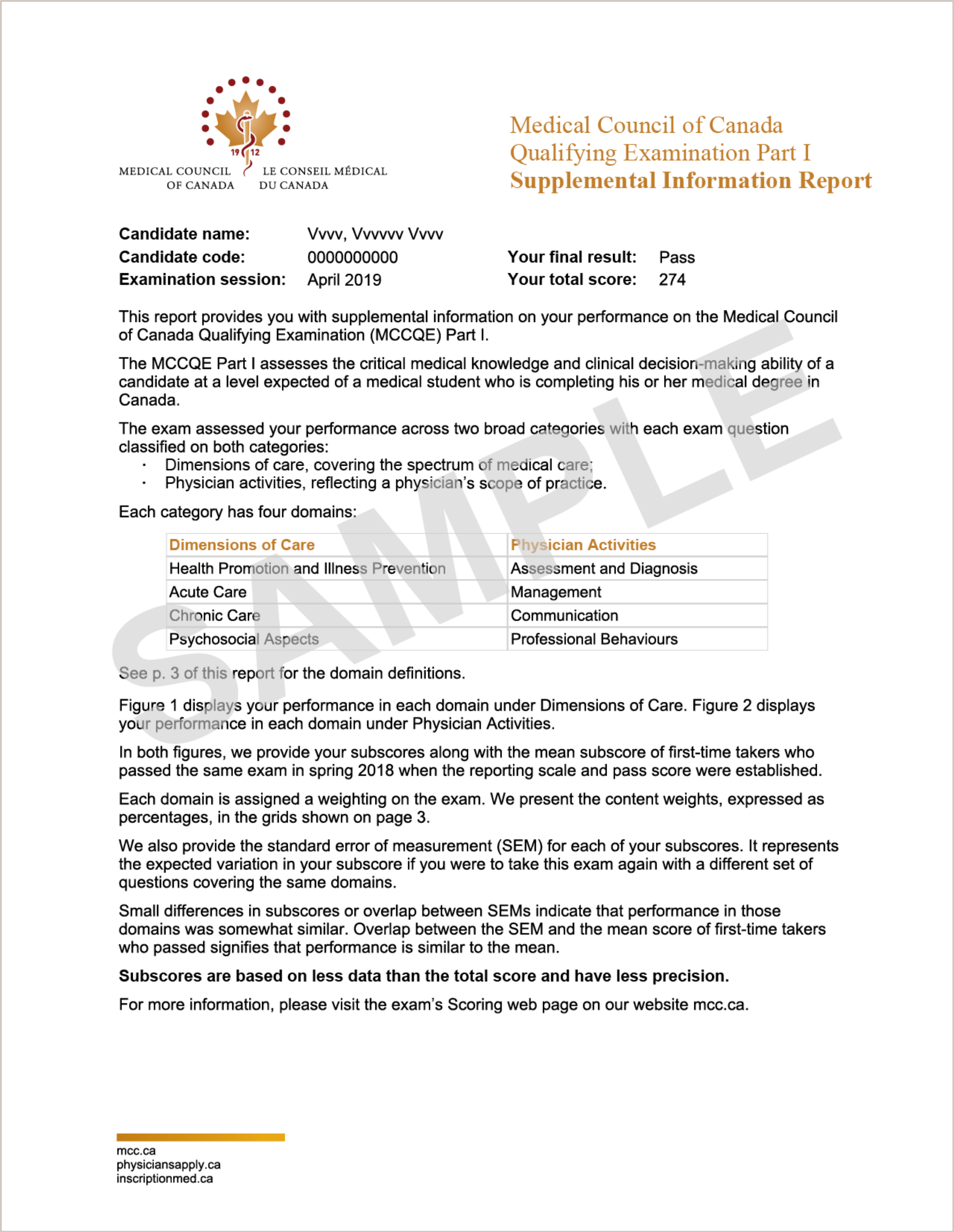
Your final result is based on your total score relative to the pass score.

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.

mcc.ca
physiciansapply.ca
inscriptionmed.ca

APPENDIX B:
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 2019	**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.

mcc.ca
physiciansapply.ca
inscriptionmed.ca

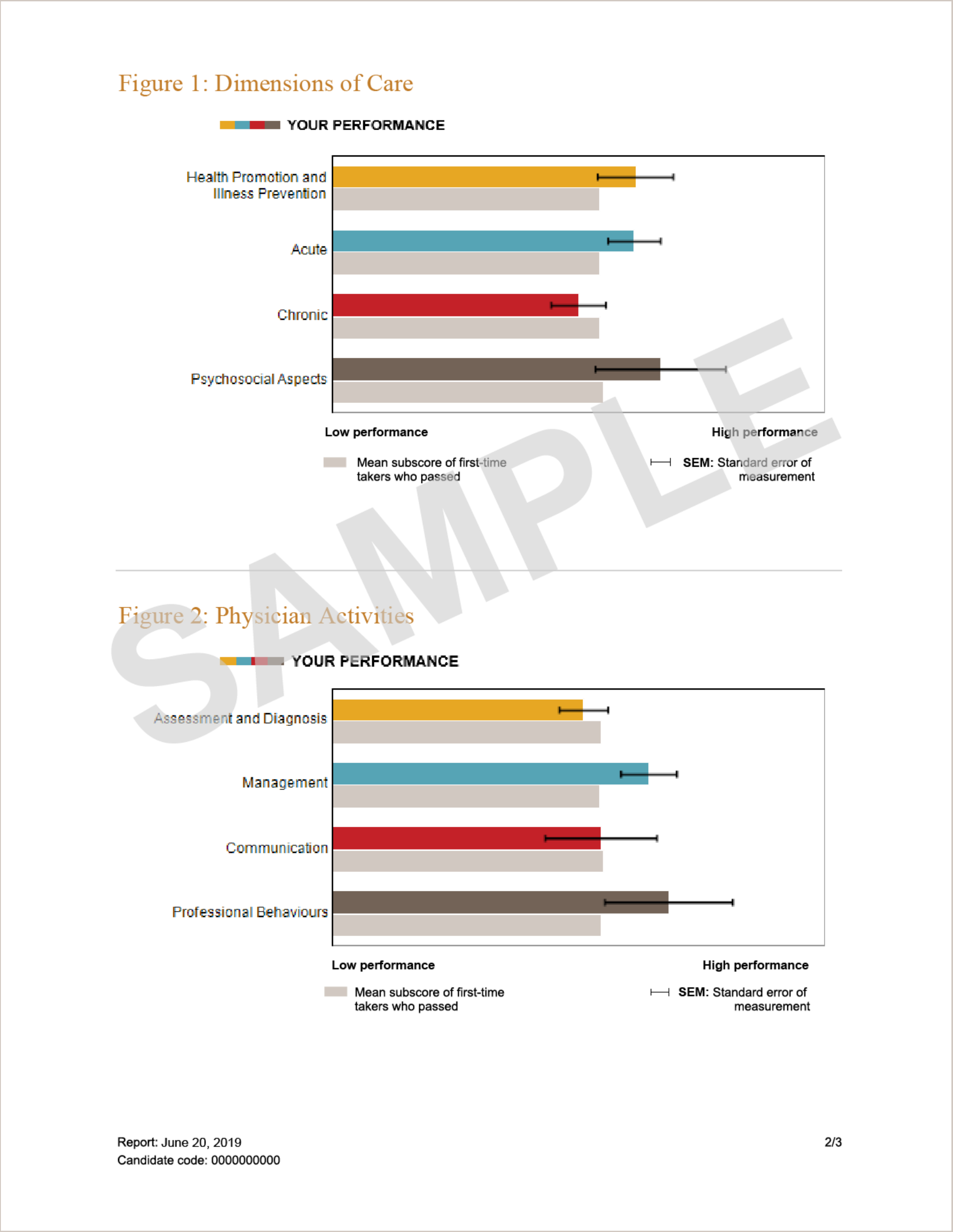


Figure 1: Dimensions of Care

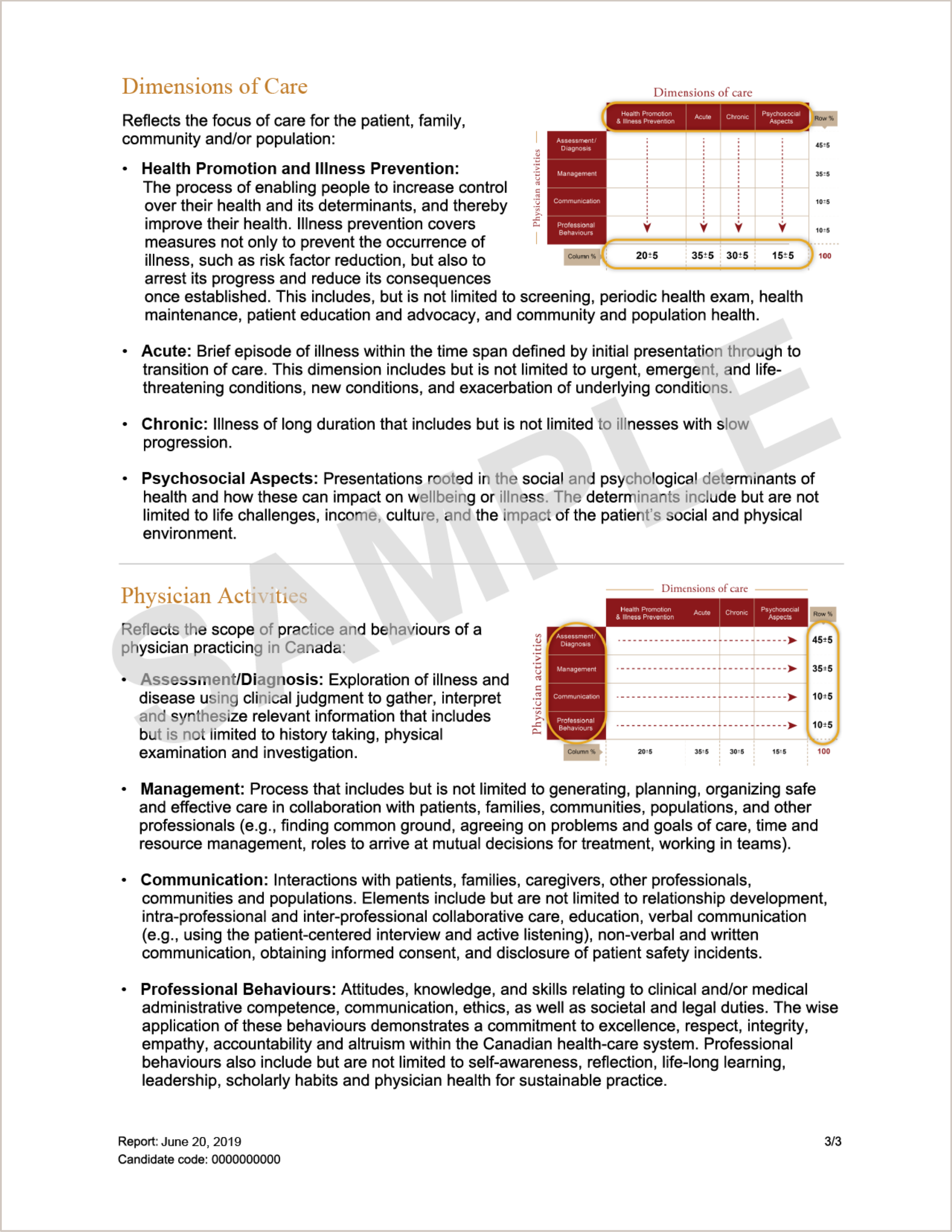
YOUR PERFORMANCE
Health Promotion and Illness Prevention
Acute
Chronic
Psychosocial Aspects
Low performance
High performance
Mean subscore of first-time takers who passed
SEM: Standard error of measurement

Figure 2: Physician Activities

YOUR PERFORMANCE
Assessment and Diagnosis
Management
Communication
Professional Behaviours
Low performance
High performance
Mean subscore of first-time takers who passed
SEM: Standard error of measurement

Report: June 20, 2019
Candidate code: 0000000000

2/3



Dimensions of Care

Dimensions of care

Physician activities

	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

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

Dimensions of care

Physician activities

	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

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.

Report: June 20, 2019
Candidate code: 0000000000

3/3

APPENDIX C: Internal structure of the MCCQE Part I

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 specialties, 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 13: Correlation matrix 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 14: Correlation matrix 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 15: Correlation matrix 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.