2024 MCCQE PART I ANNUAL TECHNICAL REPORT

MEDICAL COUNCIL OF CANADA
LE CONSEIL MÉDICAL DU CANADA

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

PREFACE

This report summarizes the fundamental psychometric characteristics, test development, test publishing, and test administration activities of the Medical Council of Canada Qualifying Examination (MCCQE) Part I. Candidate performance data on the exam sessions in April, August, and October 2024 are presented. Sections 1 to 5 describe the exam's purpose, format, content development, administration, scoring, and score reporting. These sections also provide evidence supporting score interpretation, reliability and measurement errors, and other psychometric characteristics. Section 6 summarizes candidate performances for the three sessions in 2024 and includes historical data for reference purposes. The report serves as technical documentation and reference materials for members of the Exam Oversight Committee (EOC)[1], test committee members, Medical Council of Canada (MCC) staff, the MCC Council, other interested parties, and the public.

[1] Before August 2021, the Exam Oversight Committee was known as the Central Examination Committee.

1. EXAM DEVELOPMENT

In this section, we describe the exam Blueprint, exam specifications, item development, and test assembly.

1.1 EXAM BLUEPRINT

Exam development begins with the exam Blueprint, which was approved by the MCC Council in 2014. The content specifications for the MCCQE Part I were approved by the Central Examination Committee[2] in 2016. The Blueprint addresses candidates' performance across two broad categories: Dimensions of Care and Physician Activities. There are four domains of care under each of these categories.

1. **Dimensions of Care** reflects the focus of care for the patient, family, community and/or population. Its four domains are as follows:
    a. **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 exams, health maintenance, patient education and advocacy, and community and population health.
    b. **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.
    c. **Chronic:** illness of long duration that includes but is not limited to illnesses with slow progression.
    d. **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.

[2] In 2021, the Central Examination Committee was renamed the Exam Oversight Committee.

2. **Physician Activities** reflects the scope of practice and behaviours of a physician practising in Canada and has four domains:
    a. **Assessment and 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.
    b. **Management:** a process that includes but is not limited to generating, planning and 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).
    c. **Communication:** interactions with patients, families, caregivers, other professionals, communities and populations. Elements include but are not limited to relationship development, intra- and interprofessional collaborative care, education, verbal communication (e.g., using patient-centred interviews and active listening), nonverbal and written communication, obtaining informed consent, and disclosure of patient safety incidents.
    d. **Professional behaviours:** attitudes, knowledge and skills related to clinical and/or medical administrative competence, communication and 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.

Blueprint for the MCCQE Part I

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

Table 1: Blueprint for the MCCQE Part I

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

1.2 EXAM SPECIFICATIONS

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

1.2.1 Content specifications

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

The MCQ component

The exam consists of 210 MCQs and includes pilot questions, also called pretest questions, which are scored if they perform psychometrically well. The pilot questions are not identified as pilots in the exam. Since 2020, each MCQ has a stem that includes a case description and three to five response options, of which only one is the correct answer. Candidates may select only one option in the MCQs and points are not deducted for incorrect answers. The maximum time allotted for the MCQ component is four hours.

Certain questions will include visual material, such as a photograph, a radiograph, or an electrocardiogram. If relevant to a question, normal lab values are presented directly in the question stem.

The CDM component

The exam consists of 38 CDM cases and includes pilot questions, also called pretest questions, which are scored if they perform psychometrically well. The pilot questions are not identified as pilots in the exam. Each question includes a stem that includes a case description followed by one or more options that 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
- prescribe therapy

Candidates were presented with 63 to 67 questions related to the 38 CDM cases. Responses are either in a short-menu or short-answer write-in format.

Most questions 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 harms or endangers the patient, they 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 types of questions 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 typically presented in one continuous block of time. Certain questions will include visual material, such as a photograph, a radiograph, or an electrocardiogram. If relevant to the question, normal lab values are presented directly in the question stem or in the case.

Additional content specifications

Table 2 displays the additional specifications for the MCCQE Part I.

Table 2: Additional content specifications for the MCCQE Part I

CATEGORY	DESCRIPTION
Complexity	Multiple morbidities
Age	• neonate • infant, child • adolescent • adult • adult women of childbearing age • frail elderly
Gender	Male, female, nonbinary
Special populations	Including but not limited to people who are: • Indigenous • LGBTQ2S+ • recent immigrants • living in rural areas • living with a disability • terminally ill • refugees • living with low incomes in cities • living with substance use disorders • experiencing homelessness
Setting	Included but not limited to: • rural or remote settings • long-term-care facilities • home visits

1.2.2 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 the 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% of the target.

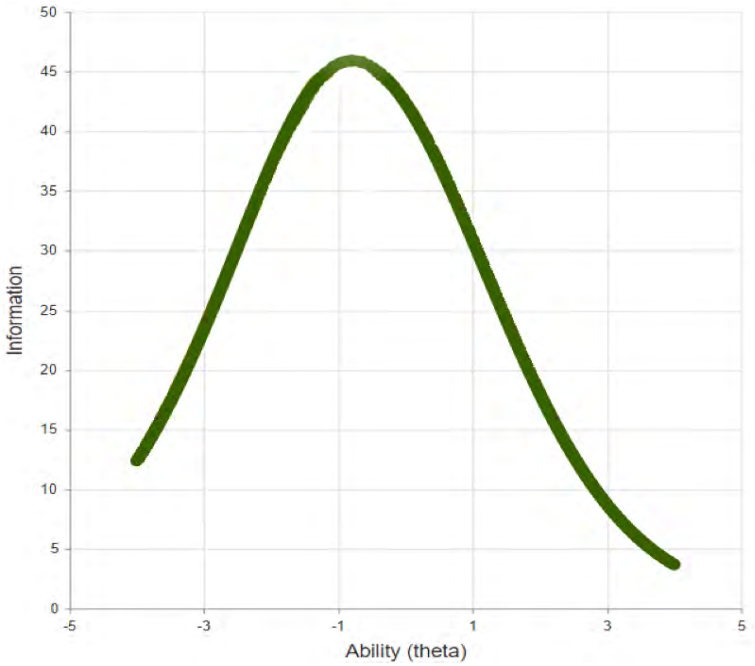


Figure 1: Target test information function

1.3 ITEM DEVELOPMENT

For MCQ content, six discipline test committees create and approve exam content. For CDM content, one multidisciplinary test committee develops exam content.

MCC's medical education advisor is an expert in medical education and assessment. The advisor attends each MCCQE Part I test committee meeting, educates item writers, instructs members on the Blueprint and MCC Objectives, supports the assessment content developers (ACDs) in identifying content gap areas, and serves as a 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 focus on specific Dimension of Care and Physician Activity topics from the Blueprint based on gaps identified in the item bank. They are also asked to consider certain test specifications, 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, and should their performance meet the statistical and content criteria, they are counted as an active item and used in scoring.

1.3.1 Test committees

Each test committee has 8 to 10 subject matter experts 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 languages (English and French) as content is produced and/or translated in both official languages.

Each test committee meets for two to three days at least once a year. 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 review, which is outlined below.

Committees develop content by following professional standards outlined in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) and the *ITC Guidelines on Test Use* (International Test Commission, 2001). These standards and guidelines include QA steps to ensure a fair assessment is delivered to the test takers.

Committee chairs and assessment content developers (ACDs) guide test committee members in the development of content where identified gaps exist in the exam Blueprint and test specifications. Item development focuses on creating items that vary in difficulty and have 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 item content using specific in-practice examples along with anticipating where errors may occur.

After the test committee vets and approves items, the English content undergoes a rigorous editorial process by the English-language editors that includes substantive editing, copy editing, and proofreading. Substantive editing includes work to improve language and structure so that content is inclusive, clear, complete, and logical. Copy editing includes fact-checking and work to correct grammar, spelling, punctuation, and mechanics while ensuring consistency and adherence to style guidelines.

Since the MCC requires the highest quality of medical translation, all translators go through a screening process to evaluate their qualifications. After translation, a team of in-house French-language editors performs an in-depth comparative read to ensure that the translation is faithful to the English version. This involves a thorough editorial and peer-review process in compliance with current French standards and MCC style guidelines. Once the edited content is approved by the ACDs and outstanding issues are resolved, the MCC conducts translation

validation sessions with the French editors and francophone physicians to make any final correction or editorial content change and ensure that the French is inclusive of regional differences. The ACDs and editors then proofread all content for a final quality control review before adding it to the pool of items available for selection and test assembly.

1.3.2 Clinical decision-making questions

The CDM test committee develops content for the CDM portion of the MCCQE Part I. This committee includes subject matter experts 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 Test committees, Section 2.3.1, and rigorous QA processes. Committee members meet twice a 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 question is the key-feature approach. This approach is based on the notion of case specificity, which means 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 three-and-a-half-hour exam, it is important to focus on the key features in the resolution of each problem, be they essential issues or specific difficulties. Test committee members think about where a minimally competent candidate would likely make 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 question 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 MCQ discipline test committees vet the content. If necessary, they send feedback and suggest revisions to the CDM test committee.

Item performance varies, and at times, items are flagged for psychometric reasons. Flagged items are 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.

1.4 TEST ASSEMBLY

Following question 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 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 the test assembly, we also consider the linking between test forms. Scores from different test forms are statistically linked through common items referred to as anchor items. Anchor items are assembled as a set of MCQs called anchor sets. 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.

ACDs 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 that is 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 do not provide answers to other test questions and that item enemies (items of similar content) are tracked to avoid appearing on the same test form.

The ACDs and psychometricians work closely to ensure the test forms are reviewed and approved by subject matter experts. Once MCC staff has vetted the forms to ensure they meet the exam specifications, a committee of experts convenes once a year to review and approve the test forms. The first step is the approval of the anchor items, and the second is the approval of the full test form of MCQs and CDMs. This is done by the test form approval committee (TFAC), which follows a thorough process to approve the test forms using the MCC's test form management system. The process for form approval is as follows:

1. The psychometrics team assembles test forms according to the exam specifications.
2. The ACDs approve the forms, exchanging any items that overlap in content or that may be an item enemy and are not yet tagged in MCC's item bank. ACDs also identify any content that may be medically inaccurate (for example, if there have been any guideline changes since item development).
3. The TFAC 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 and 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. A final review by the psychometrician and the ACDs ensures the content specifications and constraints have been respected and that 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 exam forms, where a preselected 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 a committee of physicians. Automated test assembly is used to assemble all MCCQE Part I test forms. Test forms are assembled to meet a series of content specifications, as described in Exam specifications, Section 2.2, and to be as similar as possible, both in content and difficulty. Figure 2 depicts the logic implemented to assemble a number of test forms automatically. Common items are required to establish a common scale between different test forms. The result is that scores from different test forms can be compared as they share a common scale.

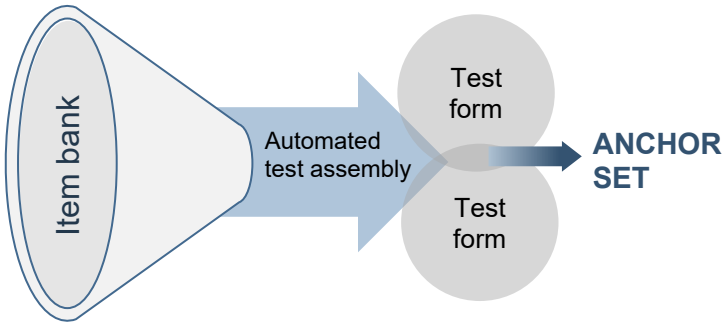


Figure 2: Automated test assembly procedure

2. EXAM ADMINISTRATION

2.1 EXAM DELIVERY

Starting in 2019, the MCCQE Part I was delivered in Canada and internationally in over 80 countries through Prometric, which has more than 20 years of experience in exam development and administration. Prometric is internationally recognized and serves professional high-stakes examination sectors. The change to Prometric ensures broader access for candidates to take the MCCQE Part I.

In 2024, the MCCQE Part I was offered during three test sessions in April, August, and October 2024. The test sessions occurred over a four- to five-week period. In both test centre and remote proctoring modalities, Prometric staff deliver the exam, follow strict exam security protocols, and monitor and support candidates' exam appointments from registration through exam completion.

The exam may be taken in English or French at any test centre and through remote proctoring; however, staff and technical support may provide service in only one language. In Canada, support in both official languages is available at the Ottawa, Montréal, and Québec City test centres.

2.2 EXAM SECURITY

The MCC takes several measures to safeguard exam security. In the COVID-19 context and with the introduction of remote proctoring, exam security remains a priority.

Exam publishing processes are well established. Guidelines and security protocols are shared and reviewed for both test centre and remote proctoring administration before each testing session. Exam results are processed in MCC's secure environment. This cycle of exam delivery offers the MCC assurances of a consistent and fair exam administration for all candidates. The MCC collaborates with interested parties 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.

Although remote proctoring poses new potential security risks (e.g., access to material that is not permitted, recording the exam), remote proctoring services are evolving and offering new AI-enabled features (e.g., facial detection) embedded in the testing software. These features are integral to mitigating security risks. They capture eye movement or the presence of additional people in the testing room, block inappropriate keystrokes, and flag security violations such as the use of smartphones and other prohibited items. Other security features include lockdown

browser functionality preventing candidates from accessing aids or capturing exam content and the use of video review if cheating is suspected.

Every site administrator and proctor is trained to recognize potential test security breaches. Training is standardized and delivered by Prometric across both modalities. Prometric conducts yearly training with all staff to communicate enhancements to security protocols and reinforce security measures. Prometric also performs regular test centre and remote proctoring audits to assess potential security gaps and ensure quick resolution.

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 indicating 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, Prometric staff may immediately terminate the candidate's exam. Prometric staff are required to produce a full candidate procedure report of all such occurrences for the MCC. The MCC also receives all candidate interaction logs to assess candidate behaviour and corroborate Prometric security concerns. 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 may take appropriate action in accordance with the MCCQE Part I Terms and Conditions candidates accepted at the time of application. In the past, the MCC has issued a Denied Standing to candidates due to irregular behaviour; consequences of this can include the following:

- The candidate may be banned from taking future MCC examinations
- The candidate's physiciansapply.ca account may be suspended
- A permanent annotation may be made on the candidate's physiciansapply.ca account
- A report may be made to medical regulatory authorities and other organizations
- Legal action may be taken against the candidate

2.3 EXAM PREPARATION

Online materials are available to assist candidates in preparing for the MCCQE Part I. These resources include the exam platform demonstration videos, sample MCQ and CDM questions, instructional videos (e.g., CDM tips, online demo), a list of resources by medical specialty area, and the *MCC Objectives*. All candidates have access to these free resources through the MCC's website.

Candidates may also test their knowledge by purchasing a full-length *Preparatory Examination* or the shorter *Preparatory Examination-Lite* through the MCC's website.

2.4 QUALITY ASSURANCE

After each exam day administration, MCC's database is updated with each candidate's response file. 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 session. A table that includes one row per item for each candidate is generated for each exam component. The tables contain the unique identifiers for candidates and items, along with the candidate answers and scores for all items. An initial round of quality assurance (QA) of the tables is performed by the psychometrician for the MCCQE Part I, including verification of completeness. Reasons for missing data are verified with the exams team. Once it is determined that the data meets the QA requirements, scoring and calibration are performed by MCC's psychometric team.

2.5 RELEASE OF RESULTS

The MCC releases candidates' results (e.g., pass or fail decision) and their 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), which provides them with information on their strengths and weaknesses by the domains in the Blueprint.

3. 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 interpretations and uses of test scores (American Educational Research Association et al., 2014). Michael T. Kane has proposed an argument-based approach to validation that involves gathering evidence to support score interpretations by establishing arguments backed by theory, empirical research, or common sense (Kane, 1990, 2013a, 2013b).

3.1 THE ARGUMENT-BASED APPROACH TO VALIDATION

According to Kane, 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 (Kane, 2013b). Any claim that certain statements about score interpretations or uses are 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 interpretations and uses of the examination. This means that scores and pass or 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. **Scoring**: Assigning scores to performance.
2. **Generalization**: Inferring expected performance across a broader universe of possible performance based on observed performance.
3. **Extrapolation**: Statements are extended to the expected performance over the domain.
4. **Implication**: Performance can also be used to make decisions about an examinee’s future.

His approach begins with an assessment of the scoring of a single observation, such as responses to exam items (Scoring), 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 (Implication).

Tables 3 to 6 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 sources of evidence (e.g., content expertise, test content, internal structure), data (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 – Scoring

Sources of evidence	Data	Warrant	Backing
Based on content expertise	Documentation, meeting notes, training slides	Items are developed to reflect relevant medical ability	• During exam content development, great care is taken to ensure the exam is relevant to medical graduates entering postgraduate training in Canada • As indicated in Exam development, Section 2, items are developed based on the exam Blueprint and content specifications defined by the CEC members • CEC (now EOC) members 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 their medical degree in Canada
Based on content expertise	Documentation, meeting notes, training slides	Proper training is offered for test developers	• Various test committees are involved in developing test items • Regular content development workshops are conducted to train test committee members to develop items that reflect the knowledge and skills emphasized in the content specifications and that meet professional test development guidelines • Guidelines have been developed for both MCQs and CDMs • The items are reviewed, edited, and finalized by test committee members, assessment content developers (ACDs), subject matter experts (SMEs), editors, and translators

Table 3: Level of inference – Scoring

Sources of evidence	Data	Warrant	Backing
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 ACDs to ensure they meet the content specifications • SMEs, ACDs, and editors review items for appropriateness of language and biased or noninclusive language or content
Based on test content	Item responses and scoring rules (MCQs and CDMs)	The answer keys are the correct answers	• Empirical evidence from item and distractor analyses is used to investigate whether the answer key is correct • For example, item-total correlations are positive for correct answers and negative for distractors • Items not meeting this expectation are identified and provided to ACDs for content review before final calibration and test scoring
Evidence of precision	Write-in item responses	Markers mark 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% agreement between markers are flagged for review • Additionally, items that have weighted kappa coefficients less than 0.61 are also flagged for review
Evidence of comparability	Candidate performance by delivery mode – 2018 to 2020	Average total scores	• Candidate average total scores in 2020 (M=247) are comparable with scores obtained in 2019 (M=252) and 2018 (M=250)
Evidence of comparability	Candidate performance by delivery mode	Average total scores	• As data from the 2020 sessions shows, Canadian medical graduates (CMGs) taking the exam for the first time in a test centre had an average performance of (M=266), which is not significantly different from CMGs who took the exam remotely (M=268) • For international medical graduates (IMGs) taking the exam for the first time, even though the difference between test centre (M=232) and remote proctoring (M=238) is statistically significant, the difference in the average total score is merely meaningful (approximately half

Table 3: Level of inference – Scoring

Sources of evidence	Data	Warrant	Backing
			of the standard error of measurement and one-fifth of standard deviation)
Evidence of comparability	Candidate performance by delivery mode	Pass rate	• As data from the 2020 sessions shows, the pass rate for CMGs taking the exam for the first time in a test centre (98%) was the same for CMGs taking the exam remotely (98%) • For IMGs taking the exam for the first time, the difference between test centre (61%) and remote proctoring (69%) is statistically significant; however, confounding factors may be interfering in these results • Confounding factors included the timing of the reopening of test centres and the registrations occurring in waves; CMGs were invited to register first (hence, reducing the spot availability for IMGs in test centres) • Also, data analyses have indicated that more prepared CMG candidates registered for the initial exam dates; this could have happened with IMG candidates as well • Another hypothesis is that extra preparation could have affected borderline IMG candidates (the impact in pass rate is more pronounced than on test scores) • The average total score for IMGs is close to the cut score (226), so small increases on their scores could cause a change in status from fail to pass

Table 4: Level of inference – Generalization

Sources of evidence	Data	Warrant	Backing
Evidence of precision	Item and test scores	The reported scores attain a level of decision accuracy and	• Using data from fiscal year 2022−2024, the decision consistency estimates have varied from 0.88 to 0.92 and the decision accuracy estimates from 0.92 to

Table 4: Level of inference – Generalization

Sources of evidence	Data	Warrant	Backing
		decision consistency that meets the target values	0.95, which indicates reliable and valid pass or fail decisions • Values were above the target value of 0.80
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 value of 0.80	• Considering the fiscal year 2022–2024 sessions, the test reliability estimates have varied from 0.90 to 0.93, indicating an adequate level of reliability of test scores, given the high-achieving characteristics of the population of examinees
Based on test content	Blueprint classification	Test forms are comparable in content	• Automated test assembly was used to assemble several fixed linear test forms, meeting almost perfectly the content specifications, as described in exam development, Section 2
Based on test content	Item parameters	Test forms are comparable in levels of difficulty	• During automated test assembly, test forms were configured to be as similar in difficulty as possible • The test information function for each of the test forms was 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	• Correlations from spring 2018 suggest that the MCCQE Part I measures an essentially single dominant underlying construct (i.e., basic medical knowledge and clinical skills that the MCCQE Part I is designed to measure) • All domains were found to be significantly and positively correlated with one another (see Appendix C) • The highest correlation was found with the total score • Correlations were also computed using the raw scores, and results support the same conclusion • This provides preliminary evidence to support the assumption of unidimensionality underlying the use of Rasch measurement models used to assemble and score the exam

Table 5: Level of inference – Extrapolation

Sources of evidence	Data	Warrant	Backing
Evidence of precision	Item and test scores	The correlation between the MCCQE Part I and NAC Examination provide some evidence of convergent validity	• The relationships between scores on the MCCQE Part I and the NAC Examination were also investigated • The NAC Examination uses an objective structured clinical examination (OSCE) format to assess the readiness of an IMG for entry into a Canadian residency program • A significant correlation ($r = .61$, $p < .0001$) was obtained between scores on the MCCQE Part I and the NAC Examination based on a sample of 1,345 candidates whose scores on both exams were matched using data from May 2018 to January 2022 for the MCCQE Part I exam and data from March 2019 to March 2020 for the NAC Examination (pre–COVID–19) • A significant correlation ($r = .51$, $p < .0001$) was obtained between scores on the MCCQE Part I and the NAC Examination based on a sample of 2,134 candidates whose scores on both exams were matched using data from May 2018 to January 2022 for the MCCQE Part I exam and data from September 2020 to October 2021 for the NAC Examination (post–COVID-19) • The correlations are 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 assess different aspects of clinical knowledge and skills

Table 6: Level of inference – Decisions

Sources of evidence	Data	Warrant	Backing

Based on standard setting	MCCQE Part I test scores and pass or fail status; subject matter expertise	Those who pass the MCCQE Part I are competent enough to practise medicine 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
			• Evidence of validity indicating that MCCQE Part I meets best practices when setting new pass scores includes:
			  ○ careful selection and training of panellists
			  ○ the methodology used on the standard-setting exercise followed best practices (Bookmark and Hofstee methods)
			  ○ feedback of the panellists' post–standard-setting exercise
			• Internal evidence included the consistency of the panellists and the convergence of results
			• Two subpanels arrived at a similar pass score independently at 95% confidence intervals constructed using standard error of judgment, which 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 in the 2018 Technical report on the standard-setting exercise for the Medical Council of Canada Qualifying Examination Part I

4. PSYCHOMETRIC ANALYSES

In 2024, the MCCQE Part I was offered during three test sessions in April, August, and October.

This section describes the psychometric analyses completed following the April exam session. We conduct item analyses, followed by item calibration, estimation of candidates' ability, scoring, standard setting and scaling (when applicable), and score reporting.

4.1 ITEM ANALYSIS: CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY

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

Classical test theory and item response theory flags

Immediately following a session, an initial item analysis is conducted using data from all first-time test takers. The initial item analysis 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, the psychometric team examines the proportion of candidates who select each option as an indicator of how well each distractor (an incorrect response) is functioning. The investigation of how well each distractor performs 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 select 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) 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. Candidates' average abilities are expected to advance across categories for CDM items. That is, a score of 2 on an item requires a 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.

CDM write-in items that display less than 90% 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 beyond chance agreement (Cohen, 1979), as it is expected that scores assigned by two markers would yield highly comparable results.

Content experts review items flagged. 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%
- Low correlation value for the correct answer: < 0.05
- High correlation value for distractor: > 0.05 and N > 10
- Top 20% 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 item response theory 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 initial item analysis and determined to be flawed after review are removed from further analyses with the subject matter experts' approval. The fall sessions are processed using the same item difficulty estimates from spring so that scores are on the same scale and thus comparable.

4.2 ITEM CALIBRATION

Previous research studies have established that simpler models, such as Rasch measurement models, yield results consistent with those from more elaborate models, such as the two-parameter item response theory logistic model (De Champlain et al., 2016; Morin et al., 2014). Starting with the spring 2015 administration, the dichotomous and partial credit Rasch measurement models were applied using Winsteps to the MCCQE Part I for item calibration and scoring (Linacre, 2015). This transition has allowed the implementation of a unified item response theory model to estimate all MCQ and CDM dichotomous and polytomous items and establish candidate abilities by considering all items together (MCQs and CDMs).

The probability of a correct response on an item (P_i) is modelled as a logistic function of the difference between a person's ability 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 using sequential integer (categorical) scores.

For the 2024 MCCQE Part I, items were recalibrated maintaining the scale established in 2018. Data from first-time Canadian medical graduate test takers was used for this process. First, the parameters for all the active 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 subject matter experts (SMEs). The decision to retain or remove those items from scoring was made. After SMEs review all flagged items (in step 1) and decide which items to remove from scoring and calibration, items are recalibrated excluding those items. A final set of calibrated items are then ready to estimate candidates' abilities.

4.3 ESTIMATING CANDIDATE ABILITY

After SMEs vet items, item parameters are used to estimate the ability of all candidates. Item parameters are fixed in the estimation process, and only the level of candidate ability is estimated.

The MCC uses the partial credit model to score candidates' exam responses (Masters, 1982). A candidate's ability and total score on the MCCQE Part I are derived from their combined performance on the MCQ and CDM components. While raw score data (scores of 1 point or zero points) 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 consider the difficulty level of the items encountered in any test 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 types of statistical software. Results are compared, and after a 100% match, they are reviewed by the psychometrician to ensure reasonableness.

Physician markers mark CDM write-in items using MCC-developed software. Physician markers are presented with CDM cases, items, key features, and scoring keys. Before the answers are presented, the software 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. Each item is marked independently by two physician markers; if discrepancies are detected, the issue is resolved by a third marker.

The software 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 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, considering all other constraints, such as exceeding the maximum number of answers allowed. The process of attributing scores to the CDM write-in items is similar to the MCQ and CDM short-menu items described above. In other words, it goes through the same rigorous QA process.

All MCQs are dichotomously scored as they all have one correct answer. Sometimes, CDM items can also be dichotomously scored. For polytomous CDM items involving more than one correct answer, successive integer scores are assigned, called category scores. For example, a candidate selecting two out of three correct answers would receive two points.

The measurement 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). It allows 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 concerning the content being assessed, whereas a candidate who obtains a score of +3.0 would demonstrate strong ability.

4.4 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[3] 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, was responsible for the quality of MCC examinations and awards final results, such as pass or fail, to candidates. The CEC approved the recommended pass score.

In the spring 2018 MCCQE Part I, a new pass score was applied to reflect this minimally acceptable level of performance. The value representing this standard was established at 0.682 on the 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.

A linear transformation is performed to establish an individual candidate's scale score. The following generic formula is applied:

[3] See the 2018 Technical report on the standard-setting exercise for the Medical Council of Canada Qualifying Examination Part I.

$$X_i' = a + bX_i$$

Where $X_i' =$ scaled score;

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

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

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

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

4.5 SCORE REPORTING

Approximately eight 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. A sample of the SOR is in Appendix A, and a sample of the SIR is in Appendix B. The SOR includes the candidate's result, total score, and 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 graphically and for feedback only and are not meant to be used by organizations for selection purposes.

If a candidate's performance has potentially been affected by procedural irregularities that occurred during an exam, the candidate may receive a No Standing, as the MCC cannot, in these cases, establish a valid pass or fail decision. In other special cases, such as candidates violating the exam's regulations (e.g., having been observed using a smartphone during the exam), the MCC may award a Denied Standing.

5. EXAM RESULTS

Candidate performance for the three sessions in 2024 is summarized in this section. When applicable, historical data from previous years are included for reference.

5.1 CANDIDATE COHORTS

The 2024 MCCQE Part I includes data from April, August, and October sessions. In 2024, the exam was administered as follows:

- a five-week session (April 17 to May 22)
- a four-week session (August 21 to September 18)
- a five-week session (October 16 to November 20)

The 7,690 candidates who challenged the exam in 2024 were educated in 141 countries, and 5,888 candidates wrote the exam in test centres in 51 countries. Table 7 summarizes the distribution of candidates across groups defined by their country of graduation and whether they were first-time or repeat test takers of the MCCQE Part I.

Table 7: MCCQE Part I group composition, 2024

	April 2024		August 2024		October 2024	
Group	No. of test takers	% of total	No. of test takers	% of total	No. of test takers	% of total
CMG[a] first-time test takers	2,390	59.1	126	7.8	105	5.2
CMG[a] repeat test takers	78	1.9	36	2.2	65	3.2
IMG[b] first-time test takers	1,116	27.6	1,088	67.1	1,321	65.1
IMG[b] repeat test takers	457	11.3	371	22.9	537	26.5
Total	4,041	100	1,621	100	2,028	100

[a]CMG: Canadian medical graduate
[b]IMG: International medical graduate
Note: Percentages do not always total 100 due to rounding

5.2 OVERALL EXAM RESULTS

Table 8 shows pass rates and basic statistics. On the score reporting scale of 100 to 400, the pass score is 226. This table does not include the two candidates who received a No Standing or Denied Standing prior to sharing the results with the EOC, but it does include 14 candidates who received a No Standing or Denied Standing after sending the results to the EOC.

Table 8: MCCQE Part I results, 2024

		April 2024	August 2024	October 2024
Canadian medical graduate (CMG) first-time test takers	No. of test takers	2,390	126	105
	Mean score	258.0	256.9	254.3
	Standard deviation	20.8	19.9	23.2
	Min. score	186	212	199
	Max. score	328	322	319
	Pass rate (%)	**94.1**	**92.9**	**87.6**
CMG repeat test takers	No. of test takers	78	36	65
	Mean score	238.6	241.9	238.7
	Standard deviation	17.4	15.8	12.6
	Min. score	197	190	212
	Max. score	281	271	273
	Pass rate (%)	**75.6**	**88.9**	**87.7**
International medical graduate (IMG) first-time test takers	No. of test takers	1,115	1,088	1,320
	Mean score	222.6	230.3	219.2
	Standard deviation	36.7	34.3	35.1
	Min. score	100	111	100
	Max. score	322	308	320
	Pass rate (%)	**52.3**	**60.7**	**46.1**
IMG repeat test takers	No. of test takers	457	371	537
	Mean score	216.7	220.8	217.5
	Standard deviation	24.2	27.1	24.7
	Min. score	115	103	104
	Max. score	290	319	288
	Pass rate (%)	**40.5**	**43.4**	**40.6**
All candidates	No. of test takers	4,040	1,621	2,027
	Mean score	243.2	230.5	221.2
	Standard deviation	32.0	32.8	32.7
	Min. score	100	103	100
	Max. score	328	322	320
	Pass rate (%)	**76.1**	**59.8**	**48.1**

Figure 3 displays the total score distribution on the reported score scale for all candidates in the three sessions and total. Overall, the total score performance of the April cohort was better than the other three cohorts.

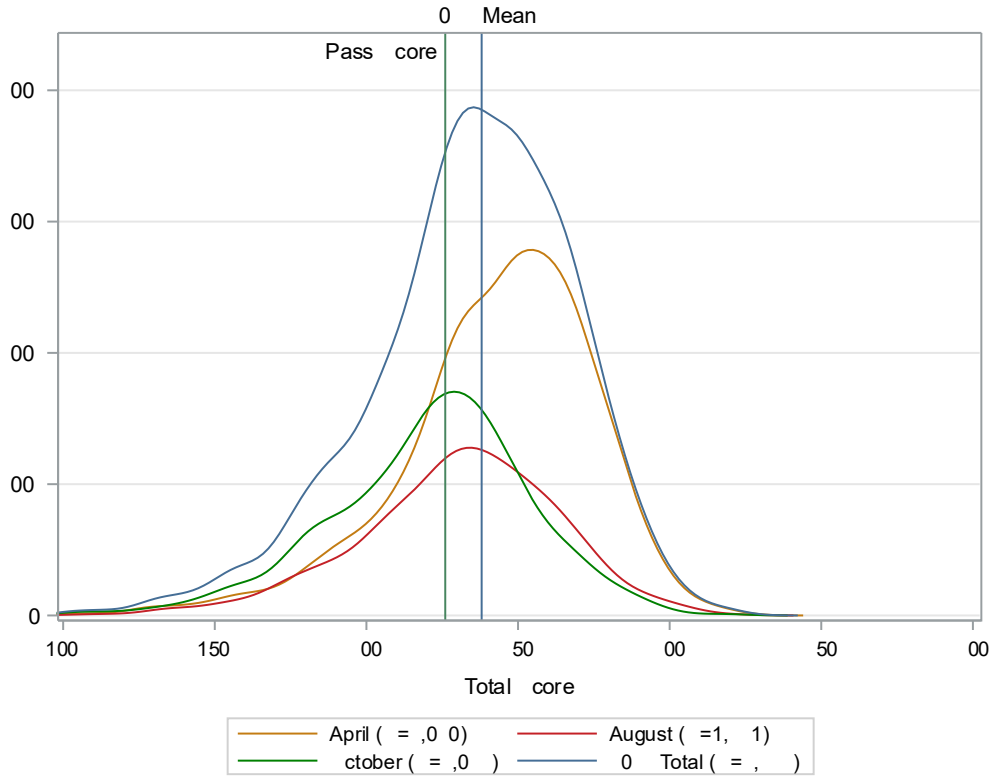


Figure 3: MCCQE Part I total score distributions, 2024

5.3 RELIABILITY OF EXAM SCORES AND CLASSIFICATION DECISIONS

In the context of this high-stakes exam, the accuracy of pass or fail decisions is of the utmost importance. Decision consistency and decision accuracy can be estimated using the Livingston and Lewis procedure (Livingston & Lewis, 1995), which is used by many high-stakes testing programs. Decision consistency is an estimate of the agreement between pass or fail final decisions on potential parallel forms of the exam. Decision accuracy is the estimate of the agreement between the pass or fail decisions based on observed exam scores and those that would be based on their true score (e.g., 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 three 2024 sessions indicate reliable and valid pass or fail decisions based on MCCQE Part I scores. Table 9 is based on data from 4,040 candidates in the April 2024 session, 1,621 in the August 2024 session, and 2,027 in the October 2024 session.

Table 9: Reliability estimates, standard errors of measurement, decision consistency and decision accuracy indices for each MCCQE Part I session, 2024

	April	August	October
Reliability estimate[a]	0.93	0.93	0.93
Average standard error of measurement (SEM) (total score)	8.4	8.2	8.1
Decision consistency	**0.92**	**0.89**	**0.89**
False positive	0.04	0.05	0.06
False negative	0.04	0.05	0.06
Decision accuracy	**0.95**	**0.93**	**0.92**
False positive	0.02	0.04	0.04
False negative	0.03	0.04	0.04

[a]Person (test) reliability from the Rasch model

5.4 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 Supplemental Information Report is designed to provide subscore information at the candidate level.

Domain subscore information for all candidates in the 2024 sessions is provided below. The range of domain subscores is presented graphically in Figures 4 to 6. The graphs show the domain subscore for each of the eight domains. The boxes for each domain indicate the range of scores for 50% of the candidates' domain subscores. The vertical line represents the median or 50th percentile subscore. The remaining 50% of domain subscores are shown to the right or the left of the box as a line (25% to the right and 25% to the left).

The legend for each of the subscores displayed in Figures 4 to 6 is as follows:

Dimensions of care	Physician activities
HEALTHP = Health promotion and illness prevention	PSYCHS = Psychosocial aspects
ACUTE = Acute	MGMT = Management
CHRONIC = Chronic	COMM = Communication
PSYCHS = Psychosocial aspects	PROFB = Professional behaviours

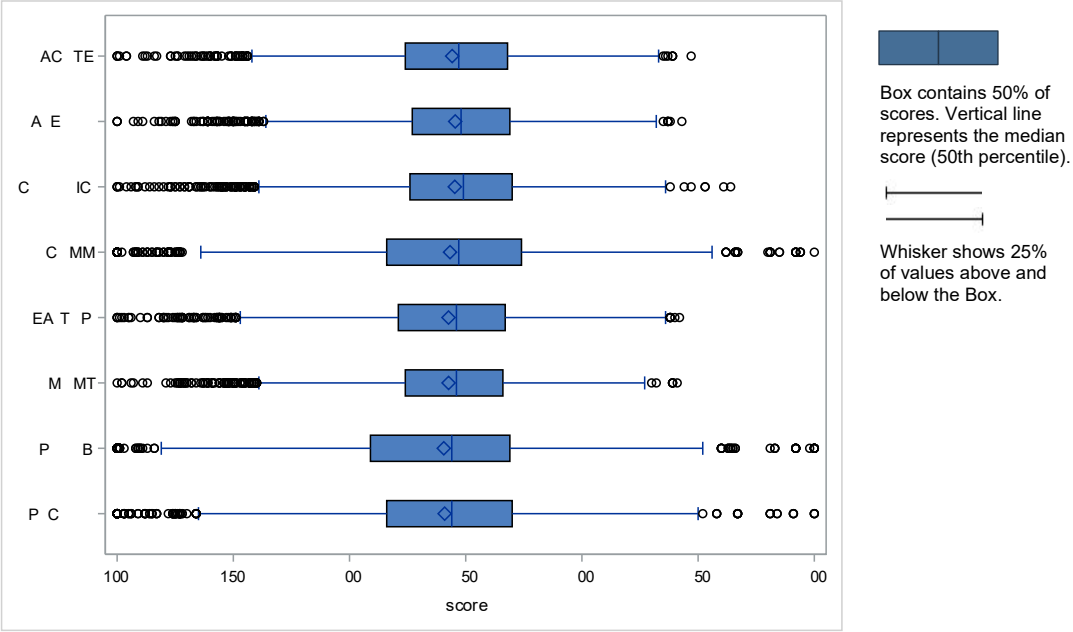


Figure 4: Domain subscore for the MCCQE Part I, April 2024 session

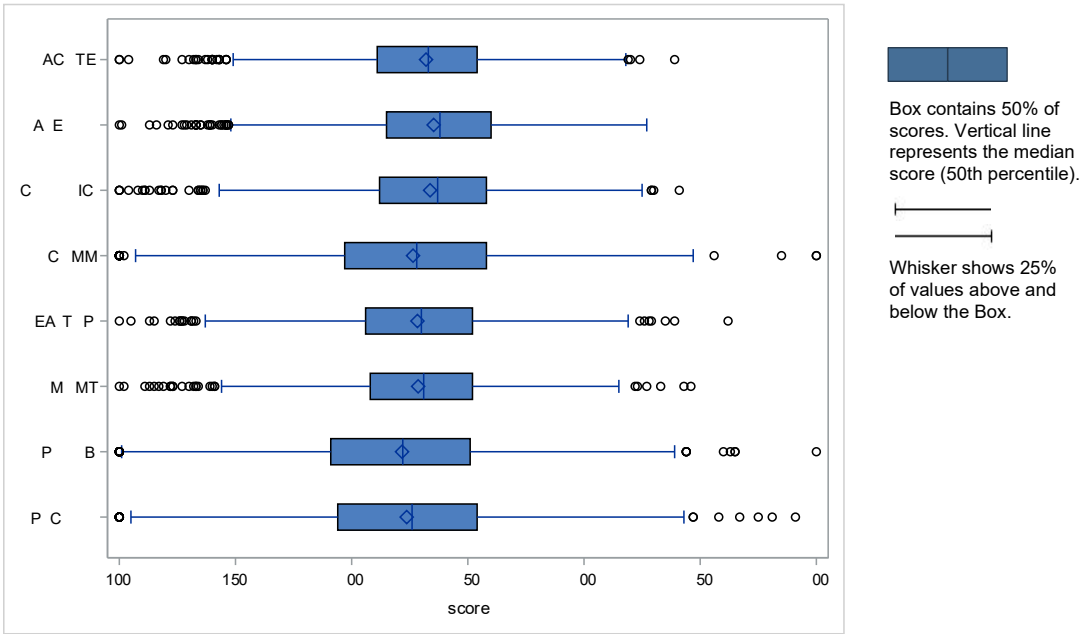


Figure 5: Domain subscore for the MCCQE Part I, August 2024 session

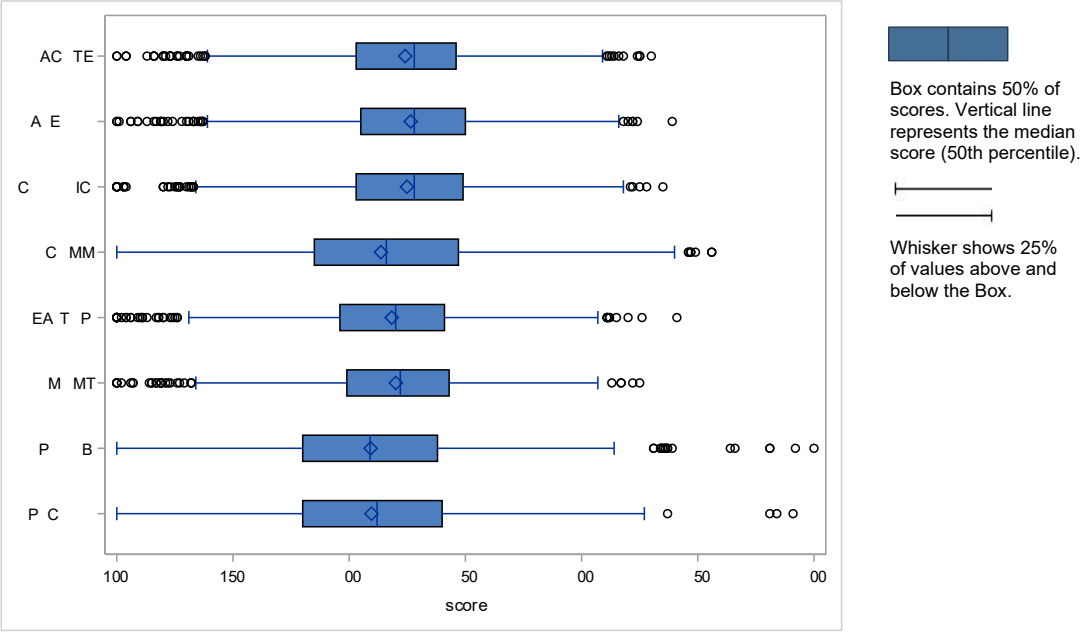


Figure 6: Domain subscore for the MCCQE Part I, October 2024 session

5.5 HISTORICAL PASS RATES

Historical pass rates are presented in this section. Table 10 shows the pass rates for 2020 to 2024 by candidate group.

Table 10: MCCQE Part I pass rates, April 2020 to October 2024

	2020–2021		2021–2022		2022–2023		2023–2024		2024	
	No. of test takers	Pass rate (%)	No. of test takers	Pass rate (%)	No. of test takers	Pass rate (%)	No. of test takers	Pass rate (%)	No. of test takers	Pass rate (%)
CMG[a] first-time test takers	2,906	98	2,919	96	2,931	93	2,979	94	2,621	94
CMG repeat takers	86	86	87	78	173	80	180	79	179	83
IMG[b] first-time test takers	2,711	64	3,140	57	2,936	59	3,266	57	3,523	53
IMG repeat takers	1,026	52	1,203	44	1,229	39	1,212	40	1,365	41
TOTAL	6,729	77	7,349	71	7,269	70	7,637	69	7,688	65

[a] CMG: Canadian medical graduate
[b] IMG: International medical graduate

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.

Cook, D.A., Brydges R., Ginsburg S., & Hatala R. (2015). *A contemporary approach to validity arguments*: A practical guide to Kane's framework. Medical Education, 49(6):560–75.

De Champlain, A.F., 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*. Journal of Educational Evaluation for Health Professions, 13, 6.

Frank, J.R., Snell, L., & Sherbino, J. (2015). *CanMEDS 2015 physician competency framework*. Royal College of Physicians and Surgeons of Canada.

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*. American College 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].

Linacre, J.M. (2016). Winsteps Rasch measurement computer program user's guide. *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.

Masters, G.N. (1982). *A Rasch model for partial credit scoring*. Psychometrika, 47, 149–174.

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

APPENDIX A:
MCCQE PART I STATEMENT OF RESULTS SAMPLE

Medical Council of Canada
Qualifying Examination Part I
Statement of Results

Candidate name:	Xxxxx, Yyyyyyyyy		
Candidate code:	0123456789		
Examination session:	April 2024	**Your final result:**	Pass
Pass score:	226	**Your total score:**	328

July 19, 2024

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 SAMPLE

MEDICAL COUNCIL OF CANADA / LE CONSEIL MÉDICAL DU CANADA

Medical Council of Canada
Qualifying Examination Part I
Supplemental Information Report

Candidate name: Xxxxx, Yyyyyyyyy
Candidate code: 0123456789
Examination session: April 2024
Your final result: Pass
Your total score: 328

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

The MCCQE Part I assesses the critical medical knowledge and clinical decision-making ability of a candidate at a level expected of a medical student who is completing his or her medical degree in Canada.

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

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

Each category has four domains:

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

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

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

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

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

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

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

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

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

Figure 1: Dimensions of Care

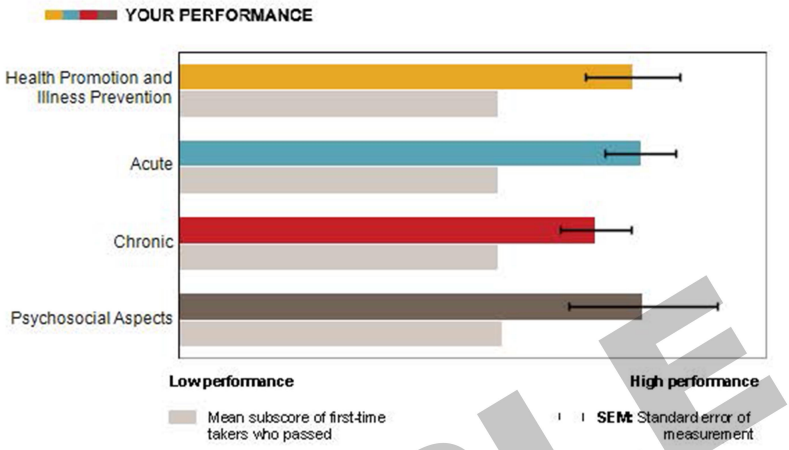
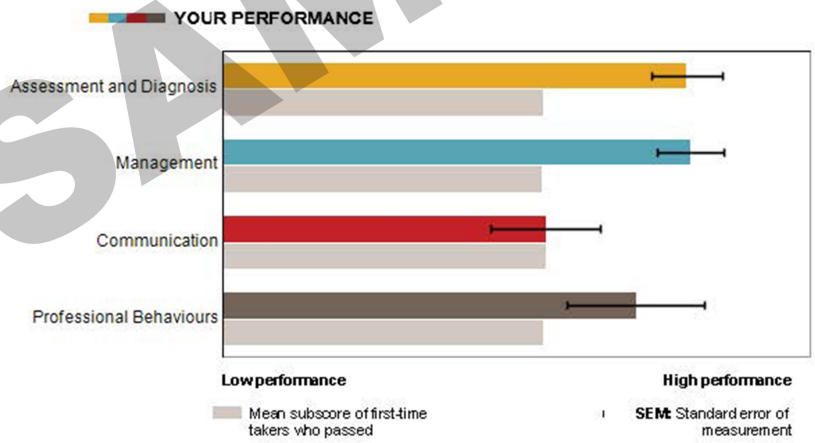


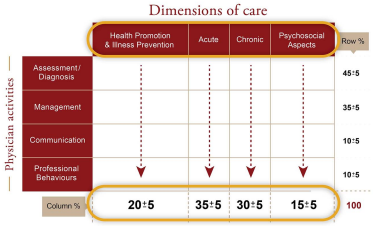
Figure 2: Physician Activities



Report: July 19, 2024
Candidate code: 0123456789

2/3

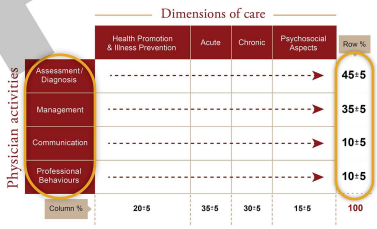
Dimensions of Care



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

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

Physician Activities



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

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

APPENDIX 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, objectives, structure, and alignment with the requirements of MCC's major partners. 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 interested parties. 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 multiple-choice question (MCQ) and clinical decision-making (CDM) components of the MCC Qualifying Examination (MCCQE) Part I.

With the Blueprint, the MCC can assess fundamental core competencies required of physicians practising in Canada at various points along their careers, regardless of specialties. It 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 evaluating the correlations among the Blueprint subscores. Correlating the two categories (and their embedded domains) can help 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 session.

Table 11: 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 and illness prevention	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 12: Correlation matrix among subscores in the four domains of Physician Activities and total scores

	Total Score	Assessment and Diagnosis	Management	Communication	Professional Behaviours
Total score	1				
Assessment and 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 13: Correlation matrix among subscores in Physician Activities and in Dimensions of Care

	Health Promotion and Illness Prevention	Acute	Chronic	Psychosocial Aspects
Assessment and 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 11 to 13, all subscores classified by either Dimensions of Care or Physician Activities were significantly and positively correlated.

This provides preliminary evidence to support the assumption of unidimensionality underlying the Rasch measurement models used to assemble and score the exam. Correlations were also computed using the raw scores, and the results supported the same unidimensionality conclusion.