

Medical Council
of Canada
Evaluating
Examination
(MCCEE)

2017 MCCEE Annual Technical Report



MEDICAL COUNCIL
OF CANADA

LE CONSEIL MÉDICAL
DU CANADA

TABLE OF CONTENTS

| | |
|---|-----------|
| PREFACE | 4 |
| SECTION 1: PURPOSE OF THE MCCEE | 4 |
| SECTION 2: EXAM DEVELOPMENT | 5 |
| 2.1 Exam format | 5 |
| 2.2 Exam specifications..... | 5 |
| 2.3 Item development..... | 7 |
| 2.4 Establishing operational item pools..... | 8 |
| SECTION 3: EXAM ADMINISTRATION | 9 |
| 3.1 Exam delivery and exam centres | 9 |
| 3.2 Exam security | 9 |
| 3.3 Exam preparation | 11 |
| 3.4 Scoring and quality control | 11 |
| 3.5 Release of results..... | 12 |
| SECTION 4: VALIDITY | 12 |
| 4.1 Evidence based on exam content | 12 |
| 4.2 Evidence based on internal structure | 13 |
| 4.3 Evidence based on relations to other variables | 14 |
| 4.4 Minimizing construct-irrelevant factors | 15 |
| SECTION 5: PSYCHOMETRIC ANALYSIS | 16 |
| 5.1 Item analysis..... | 16 |
| 5.2 Item bank calibration | 17 |
| 5.3 LOFT delivery | 18 |
| 5.4 Standard setting | 20 |
| 5.5 Scoring and score reporting | 20 |
| SECTION 6: EXAM RESULTS | 21 |
| 6.1 Candidate cohorts | 21 |
| 6.2 Overall Exam Results..... | 23 |
| 6.3 Reliability of exam scores and classification decisions | 24 |
| 6.4 Candidate performance by sub-category | 26 |
| 6.5 Exam results by candidate group | 28 |
| 6.6 Comparison of prior exam performance..... | 29 |
| 6.7 Item exposure analysis..... | 30 |
| 6.8 Candidate survey | 30 |
| REFERENCES | 32 |
| APPENDIX A: List of countries where the MCCEE is offered | 33 |
| APPENDIX B: Statement of results (SOR) example | 34 |
| APPENDIX C: Supplemental feedback report (SFR) example | 35 |

LIST OF TABLES AND FIGURES

| | | |
|------------------|--|----|
| Table 1: | Exam content specifications for the MCCEE – Number of items per health group and clinician task | 6 |
| Table 2: | Correlations among subscores in health groups (N = 3,811) | 14 |
| Table 3: | Correlations among subscores in clinician tasks (N = 3,811) | 14 |
| Table 4: | Correlations among subscores by specialty area (N = 3,811) | 14 |
| Table 5: | Correlations between scores on the MCCEE and other MCC exams | 15 |
| Table 6: | Distribution of candidates in 2017 by region | 21 |
| Table 7: | Distribution of candidates in Canadian test centres in 2017 by administration | 22 |
| Table 8: | Distribution of candidates in 2017 by group and administration | 22 |
| Table 9: | Descriptive statistics for the total score and pass rates in 2017 by administration | 23 |
| Table 10: | Estimates of decision consistency and decision accuracy in 2017 | 25 |
| Table 11: | Descriptive statistics for the total score and subscores in 2017 | 26 |
| Table 12: | Descriptive statistics and pass rates in 2017 by candidate group | 28 |
| Table 13: | Pass rates of each 2017 administration and the previous five years | 29 |
| Table 14: | Item exposure in 2017 | 30 |
| Table 15: | Candidate Survey Results (2017) | 30 |
| | | |
| Figure 1: | Exam psychometric specifications for the MCCEE – Target Test Information Function | 7 |
| Figure 2: | Total score distributions in 2017 | 23 |
| Figure 3: | Distributions of standard errors of the ability estimates for 2017 cohorts | 24 |
| Figure 4: | Subscore distributions for health groups in 2017 | 27 |
| Figure 5: | Subscore distributions for clinician tasks in 2017 | 27 |
| Figure 6: | Subscore distributions for specialty areas in 2017 | 28 |

PREFACE

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

SECTION 1: PURPOSE OF THE MCCEE

The MCCEE is a four-hour, computer-based exam offered in both English and French in over 80 countries worldwide. International medical students and American osteopathic students in the final 20 months of their program and international medical school graduates or American osteopathic graduates must take the MCCEE as a prerequisite for eligibility for the MCC Qualifying Examination (MCCQE) Part I. The MCCEE is also a prerequisite for the National Assessment Collaboration (NAC) Examination, an Objective Structured Clinical Examination (OSCE) designed to assess the readiness of an international medical graduate (IMG) for entry into residency training programs in Canada. However, starting in March 2018, a pass on the MCCEE. will no longer be an eligibility requirement to apply to the NAC exam.

The MCCEE is a screening examination that assesses the basic medical knowledge and problem solving of a candidate at a level comparable to a minimally competent medical student completing his or her medical education in Canada and about to enter supervised practice. It provides the candidate with an estimate of the probability of his or her chances of succeeding in the Canadian system.

The EECC is responsible for overseeing the MCCEE including the development of the exam, the maintenance of its content and the approval of results.

SECTION 2: EXAM DEVELOPMENT

2.1 Exam format

The MCCEE consists of 180 multiple-choice questions (MCQs) including 150 operational items¹ (scored items) and 30 pilot items (new, non-scored items pretested for future use). The items cover child health, maternal health, adult health (including gynecology, medicine and surgery), mental health and population health and ethics. A number of items in the exam also focus on general practice.

Each item lists five possible answers of which only one is correct. The MCCEE is administered using a computer-based, Linear-On-the-Fly-Test (LOFT) model and is delivered securely by Prometric, a test delivery provider. With the LOFT design, a unique exam form is assembled in real-time whereby items are selected from a large pool of operational items based on exam specifications, as described in the following section, each time a candidate takes the exam. More detailed explanations of the LOFT design are provided in Section 5.3.

2.2 Exam specifications

The exam specifications for the MCCEE define the content and psychometric specifications for each exam. The content specifications include the content domains to be tested, a sampling plan for the content domains (the proportion of items per content area) and total exam length (total number of items). The psychometric specifications include the desired psychometric properties of the items (number of items for each level of difficulty), target standard error of ability estimates and an overall target test information function for each exam. The exam specifications were created and adopted by the EECC between 2008 and 2009 during a one-week retreat of the EECC and the Australian Medical Council (AMC). During the workshop, the EECC and the AMC devised a realistic representation (percentage-wise) by health group and clinician task, of what physicians would encounter in their practice on a daily basis, which, in turn, became the exam specifications.

Table 1 outlines the content specifications, including the definitions of the various health groups and clinician tasks.

¹ The term “question” and “item” are used interchangeably in this report and should be treated synonymously.

**Table 1: Exam content specifications for the MCCEE –
Number of items per health group and clinician task**

| | HEALTH GROUP | | | | | TOTAL |
|--|--------------|-----------------|--------------|---------------|----------------------------|------------|
| | Child Health | Maternal Health | Adult Health | Mental Health | Population Health & Ethics | |
| CLINICIAN TASK | | | | | | |
| <i>Data gathering</i> | 7 | 4 | 20 | 7 | | 45 |
| <i>Data interpretation & synthesis</i> | 9 | 4 | 26 | 9 | 13 | 54 |
| <i>Management</i> | 9 | 5 | 28 | 9 | - | 51 |
| TOTAL | 25 | 13 | 74 | 25 | 13 | 150 |

HEALTH GROUPS:

Child Health

Issues particular to individuals up to the end of adolescence.

Maternal Health

Issues related to pregnancy and childbirth.

Adult Health

Issues specific to individuals after the end of adolescence in medicine, surgery and gynecology.

Mental Health

Biopsychosocial/cognitive issues related to mental health in all age groups.

Population Health and Ethics

Issues related to groups and ethical behaviour. This includes population issues such as immunization, disease outbreak management, population screening and surveillance, health promotion strategies, epidemiology and relevant statistics. Ethical issues include boundary issues, impairment of doctors and informed consent.

CLINICIAN TASKS

Data gathering

History taking, mental status examination, physical examination, laboratory testing, other modalities (e.g., imaging, EKG, EEG, etc.).

Data interpretation and synthesis

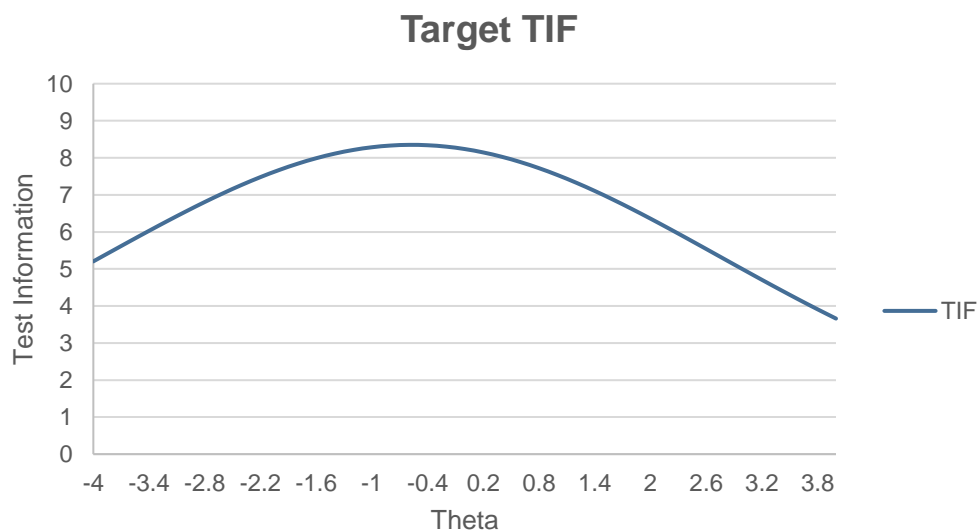
Interpretation and synthesis of gathered data. Problem identification, setting priorities, risk stratification and the formulation of differential and specific diagnoses.

Management

Education and health promotion, counselling, psychotherapy, drug and non-drug therapy (includes fluid and electrolyte therapy etc.), surgical interventions, radiological interventions, cessation of therapy, rehabilitation, palliative care, interdisciplinary management, family and community care.

The psychometric specifications set out the desired statistical properties for the exam and the items in each exam form. They include the target Test Information Function (TIF) across the ability range as indicated in Figure 1. For the MCCEE, each candidate receives a different exam form. The target TIF is used to balance multiple forms and to ensure that precision of measurement across the ability scale is highly comparable from one test form to another. The MCCEE is designed to provide maximum information (precision of measurement or reliability) and minimum error near the cut score ($\theta = -.490$) to achieve optimal precision at the cut score and consequently, maximize pass or fail decision consistency and accuracy. Section 5.4 explains how the cut score is established.

Figure 1: Exam psychometric specifications for the MCCEE – Target Test Information Function



2.3 Item development

The MCCEE items are developed by six specialty area specific test committees: Medicine, Obstetrics & Gynecology (OBGYN), Pediatrics, Population Health and Ethics, Psychiatry and Surgery. Each committee comprises six to eight physicians from across Canada who are subject matter experts (SMEs) in their fields and experienced in medical education and assessment. SMEs are recommended by test committee members or by the MCC Selection Committee. The MCC Selection Committee presents the test committee membership to Council at the Annual Meeting for approval.

Test committees include representation from both official language groups (English and French) and geographic representation from across Canada. At least two family physicians are represented on each committee and membership is diverse, representing both rural and

urban experiences. When possible, selecting physicians from a variety of teaching programs and medical education interests is preferable.

Training is provided to item writers. Training resources for test committee members is available on the MCC's website, in addition to training that occurs during content development workshops.

Test items are developed in accordance with professional standards for item development and review as outlined by the American Education Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (AERA, APA, NCME; 2014) and following the high-level process outlined here. Each test committee convenes once per year in Ottawa at which time MCQs are written, classified, peer-reviewed and approved for piloting. MCC's Test Development Officer (TDO), in conjunction with the Test Committee Chair, offers guidance to test committee members as they develop items to address known content gaps. Development is focused on creating items with a range of difficulty levels, updating items to reflect new medical terminology (DSM-5, new medical practice and treatments, etc.), adding items required to meet test specifications and/or creating items that fill content gaps in the item bank.

All new and approved items from each test committee are reviewed and approved for piloting by the EECC (a multi-disciplinary committee composed of the chairs and vice-chairs of the six specialty area test committees). The EECC conducts an overall review of items for bias and sensitivity to ensure the test items and stimuli are fair for the candidates. Once all content has been approved, all items are sent for editorial review by MCC's Examination Content Editors prior to being sent for translation. Linn (2006) states, "Even skilled and experienced item writers sometimes produce flawed items that are ambiguous, have no correct answer, or are unintentionally offensive to some groups of test takers. Hence, it is critical that items be subjected to critical review and editing prior to inclusion in a test" (p. 32). Approved pilot items are then included on a pilot test form. Newly-created items are piloted before they are used as operational items on any examination form. Each pilot form contains 30 items, with five items from each specialty area.

Though all pilot items are embedded in the operational exam, they do not count toward the candidate's final score. Pilot items are analyzed and calibrated when enough data has been collected. Items that do not perform as expected are returned to the test committee for review/revision and are later re-piloted. Approximately 180 to 250 items are piloted across the five MCCEE administrations each year. Pilot items that meet content and psychometric criteria are added to the item bank for future operational use.

2.4 Establishing operational item pools

Each year, the EECC meets to establish and approve a pool of 1,500 operational items drawn from the MCCEE item bank (see Section 5.2). The TDO, using the item pool assembly tool in the item bank, selects all items not used in the prior pool and adds them to a pool of

available items for establishing a new operational item pool. The item pool assembly tool then establishes a new operational pool using fixed content and psychometric constraints based on the exam specifications. Some manual processing is performed to meet the content and psychometric specifications; the goal is to create a unique pool each time, though there is some occasional overlap across pools. This process occurs 14-16 months before the item pool is used.

The EECC performs a final quality assurance check of all item content and sees to it that the scoring key is correct. If an item is no longer valid, a replacement item is chosen from a pre-selected set of potential replacements from the item bank. Each replacement item must meet the specifications of the discarded item with respect to content area (health group, clinician task) and difficulty level. Exam forms are assembled to meet test specifications as items are drawn from the final approved operational item pool.

SECTION 3: EXAM ADMINISTRATION

3.1 Exam delivery and exam centres

The MCCEE is offered in January, March, May, September and October-November of each year. Each session consists of a two- to three-week testing window. Prometric is the vendor sourced by the MCC to deliver the MCCEE globally.

Eligible candidates are able to self-schedule their exam through the Prometric website. There are more than 500 Prometric test centres in approximately 80 countries. Scheduling is done on a first come, first-served basis.

A list of countries where the MCCEE is offered appears in **Appendix A**.

3.2 Exam security

“Security is a major concern for test administration” (Downing, 2006, p.1). The MCC has a comprehensive approach to address exam security. This includes; registration, content development, content transfer, test publishing and delivery, exam sites, the secure transfer of results back to the MCC for scoring, and results analysis. This “chain of security” is required during test production and widens even more during larger-scale test administrations (Downing, 2006, p.15).

Registration:

In physiciansapply.ca, MCC's online registration portal, only authenticated and eligible candidates are permitted to register for the exam. Once registered, candidates receive an "authorized to test" (ATT) identification number that is required by Prometric to schedule an exam. These initial registration processes validate that only approved test takers can register and attempt an exam.

Content development:

The MCC communicates regularly with subject matter experts (SMEs) the importance and priority of exam content security. All SMEs are required to sign a confidentiality and conflict of interest agreement with the MCC. This is also a requirement for all MCC staff.

Examination content is developed during on-site meetings at the MCC's headquarters through a secure item banking software developed and stored internally. Content writers, when required to work remotely, log in to the MCC servers using a secure two-step authentication process.

Secure test publishing processes and protocols have been well established with Prometric and test centre guidelines (test delivery) and are reviewed with them prior to each testing window to ensure that results are processed in a secure environment.

Content transfer:

Content is transferred between the MCC and Prometric using a secure File Transfer Protocol (FTP). The content resides in the secure Prometric environment while staff run necessary analysis testing the delivery algorithms and reporting on any blueprint inconsistencies revealed during this simulation.

The MCC staff log into a secure Prometric system to review the content for any errors or formatting issues. When all content issues are resolved and blueprint simulations validated, the examination is ready for delivery on Prometric's secure platform.

Test publishing and delivery:

Test publishing processes, using the LOFT delivery method, limits the exposure of the entire MCCEE item bank. From a test security perspective, this delivery method administers only a portion of the pre-selected item pool and a unique form to each candidate. Even if content is shared amongst candidates, the likelihood of a test taker seeing the same item is significantly reduced. The LOFT pool is typically updated yearly.

Exam sites:

The uniform design of the Prometric labs worldwide delivers a consistent exam environment where security is of highest priority. Upon arrival, each candidate is asked to secure their personal belongings, including smartphones and other transmitting devices, in a locker prior to entering the testing room. All candidates are required to provide government issued

identification to confirm their identity. As the candidate is checked into the Prometric registration system, site staff is required to confirm that the presented candidate matches the photo identification supplied by the MCC. All candidates are then screened for electronic devices, either with a physical wand or by passing through a full-body scanner. Additionally, candidates are monitored throughout the exam by site staff passing through the exam centre and through video surveillance.

Proctors at every Prometric testing center have been professionally trained to identify potential test security breaches and each location is monitored with advanced security equipment and subject to multiple, random security audits.

During an active examination session, daily Centre Procedure Reports (CPRs) are sent to the MCC for evaluation and investigation, along with the appropriate investigative materials available from Prometric (videos, documentation from the proctor, etc.).

Exam results and analysis:

At the conclusion of an examination, candidate results are transferred to the MCC via secure FTP and processed in the MCC's secure scoring environment.

The MCC staff analyzes candidate performance by exam date over each testing window, searching for evidence of any content exposure and/or security breaches. In addition, the MCC monitors various social media websites in search of disclosure of test content and investigate if any security breach is identified.

3.3 Exam preparation

Online materials are available to help candidates prepare for the MCCEE. These resources include a demonstration of exam format, computer navigation, self-assessment tools, a list of reference manuals by specialty area and the MCC Objectives. Candidates can access all resources on the MCC's website at mcc.ca/examinations/mccee/preparation-resources.

3.4 Scoring and quality control

The Evaluation Bureau uses a number of technological systems and scoring applications to perform an initial quality assurance and data validation. Once it is determined the data meets the established quality assurance requirements, the final scoring is completed by the Evaluation Bureau and exam results are analyzed and summarized in a report by Psychometrics and Assessment Services (PAS).

The MCCEE results are reported on a standard score scale that ranges from 50 to 500; the pass mark was set at 250 before May 2017 and at 261 beginning with the May 2017 administration. (Please refer to Section 5.4 for details regarding standard setting.) Before scores are released, exam results are reviewed and approved by the EECC.

3.5 Release of results

Approximately six to eight weeks following the last day of the exam session, the EECC meets to review performance on the exam, address administrative issues, rule on special candidate cases and approve exam results. Starting in September 2017, the EECC has deemed exam results auto-approved if exam psychometric performance and candidate performance fall with the established parameters for auto-approval. Any special cases that require the EECC's review will continue to be brought to the EECC for discussion and decision. The MCC then grants candidates access to their final result (pass or fail) and total score through their physiciansapply.ca accounts. Shortly thereafter, each candidate has access to the statement of results (SOR), the official results document, and the supplemental feedback report (SFR), providing information on their relative strengths and weaknesses by health group, clinician task and specialty area.

Samples of an SOR and an SFR are available in **Appendix B and C**, respectively.

SECTION 4: VALIDITY

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

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

4.1 Evidence based on exam content

During the course of exam content development, care is taken to ensure the exam is relevant to undergraduate medical education (UGME) and to the requirements for entry into postgraduate training in Canada. As indicated in Section 2, the MCCEE items are developed based on exam content specifications carefully defined by the EECC members who ensure the exam content reflects the basic medical knowledge and problem solving of a candidate at

a level comparable to a minimally competent medical student completing his or her medical education in Canada and about to enter supervised practice. As the MCCEE is designed for international medical graduates (IMGs) who may be less familiar with the Canadian practice environment, particular attention is paid to ensuring the exam is free of content such as medical practice, therapeutics, and legal/ethical issues specific to Canada.

Various test committees are involved in developing test items. Regular content development workshops are conducted to train committee members on professional test development guidelines and on drafting items that reflect the knowledge and skills emphasized in the exam specifications for each content area. The draft items are reviewed, edited and finalized by test committee members, TDOs and editors. The items are initially developed in English and then translated into French by professional translators whose work is vetted by TDOs and editors. In addition, an analysis is performed after each exam administration to ensure that all exam forms assembled during an administration comply with the exam content specifications. These rigorous approaches all help ensure content validity of the MCCEE.

4.2 Evidence based on internal structure

As each candidate receives a different (but comparable) set of items, a factor analysis cannot be conducted to examine the factor structure of the exam. However, the internal structure of the MCCEE can be revealed, to some degree, through the evaluation of the correlations among subscores of health groups, clinician tasks and specialty areas. This can help one understand how closely the exam conforms to the construct of interest. Correlations among subscores were examined using the data from 3,811 candidates who took the MCCEE in 2015.

Tables 2, 3 and 4 present the correlation matrices among subscores in the five health groups, three clinician tasks and six specialty areas, respectively. The term discipline is an old classification system no longer used to assemble the MCCEE forms and are now referred to as specialty areas.

As indicated in each table, all subscores classified by either health group, clinician task or specialty area were found to be significantly, positively correlated with one another. This suggests that the MCCEE seems to measure an essentially single dominant underlying construct (basic medical knowledge and clinical skills that it is designed to measure). Furthermore, this provides some preliminary evidence to support the assumption of unidimensionality underlying the use of the item response theory (IRT) model (see Section 5) used to assemble the exam. It should be noted that the magnitude of correlations was affected by the number of items in each domain. For example, the higher correlations among the three clinician tasks were likely due to the larger number of items in these domains. Conversely, since there were fewer items in Population Health & Ethics, its correlations with other domains were affected.

Table 2: Correlations among subscores in health groups (N = 3,811)

| | Child Health | Maternal Health | Adult Health | Mental Health |
|----------------------------|--------------|-----------------|--------------|---------------|
| Maternal Health | 0.45* | | | |
| Adult Health | 0.68* | 0.52* | | |
| Mental Health | 0.53* | 0.38* | 0.60* | |
| Population Health & Ethics | 0.41* | 0.31* | 0.50* | 0.48* |

*significant at $p < 0.0001$

Table 3: Correlations among subscores in clinician tasks (N = 3,811)

| | Management | Data gathering |
|---------------------------------|------------|----------------|
| Data gathering | 0.69* | |
| Data interpretation & synthesis | 0.74* | 0.73* |

*significant at $p < 0.0001$

Table 4: Correlations among subscores by specialty area (N = 3,811)

| | Medicine | OBGYN | Pediatrics | Surgery | Psychiatry |
|----------------------------|----------|-------|------------|---------|------------|
| OBGYN | 0.56* | | | | |
| Pediatrics | 0.61* | 0.56* | | | |
| Surgery | 0.62* | 0.53* | 0.56* | | |
| Psychiatry | 0.55* | 0.51* | 0.53* | 0.48* | |
| Population Health & Ethics | 0.45* | 0.41* | 0.41* | 0.40* | 0.48* |

*significant at $p < 0.0001$

4.3 Evidence based on relations to other variables

The relationships between scores on the MCCEE, the MCCQE Part I and the NAC Examination were reviewed for convergent validity evidence. Both the MCCEE and the MCCQE Part I assess essential medical knowledge and skills at the level of new medical graduates about to enter the first year of postgraduate training. The MCCEE is a prerequisite for IMGs who wish to take the MCCQE Part I or the NAC Examination. The NAC Examination uses an OSCE format to assess the readiness of an IMG for entry into a Canadian residency program.

Correlations between scores on the MCCEE, the MCCQE Part I and the NAC examination are presented in Table 5. A significant correlation ($r=.70$, $p<.0001$) was obtained between scores on the MCCEE and the MCCQE Part I based on a sample of 2,071 candidates for whom the data between the two exams were matched. This provides evidence of high convergent validity between the two exams. A significant correlation ($r=.38$, $p<.0001$) was also obtained between scores on the MCCEE and the NAC Examination based on a sample of 1,711 candidates whose scores on both exams were matched. 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. The correlations between the MCCEE and the other two exams could have been higher if not due to range restriction on the former. Table 5 also presents disattenuated correlations between the MCCEE and the other two exams. The disattenuated correlation between two exams is based on their observed correlation adjusted for reliability of the exams and it indicates what their correlation would be after correction for measurement error.

Table 5: Correlations between scores on the MCCEE and other MCC exams

| | MCCEE | | N |
|------------------------|----------------------|---------------------------|------|
| | Observed Correlation | Disattenuated Correlation | |
| MCCQE Part I | 0.70* | 0.78* | 2071 |
| NAC Examination | 0.38* | 0.47* | 1711 |

* $p<.001$

4.4 Minimizing construct-irrelevant factors

Another way to enhance validity is through the minimization of construct-irrelevant variance (error variance unrelated to the construct measured by the exam). During development, items are reviewed by SMEs and TDOs to ensure they meet the exam specifications. SMEs and TDOs also review items for appropriateness of language and potential bias against certain language or culture groups. In addition, empirical evidence from item and distractor analysis is used to further investigate potential sources of construct irrelevance. For example, distractors with positive point-biserial correlations may indicate that an item is assessing a construct that is unrelated to the one intended to be measured. Test completion rates, candidate item response times and overall test times are also analyzed to ensure the time allotted to complete the exam is adequate and that speededness is not a factor affecting candidate performance. Through Prometric, the MCC ensures that testing conditions across all test centres are standardized so that candidates have equal opportunities to demonstrate their ability. Finally, detailed test information and links to resources are provided on the MCC's website to help candidates prepare for the exam and alleviate test anxiety.

SECTION 5: PSYCHOMETRIC ANALYSIS

5.1 Item analysis

The MCCEE items are analyzed using both Item Response Theory (IRT) and Classical Test Theory (CTT) frameworks. As described in Section 2, each exam form consists of 180 multiple-choice items including 150 scored operational items and 30 non-scored pilot items. The exam forms are assembled online in real-time by drawing items from a large, operational item pool built from the MCCEE item bank (see Sections 2.4 and 5.3). All items in the bank have been field tested and between 400 and 500 new items are created and piloted each year. Before pilot items are uploaded into the item bank, they are assessed for quality, analyzed and calibrated to the common scale of the item bank (see Section 5.2). Item analysis involves computing a set of statistics based on both IRT and CTT. These statistics provide information about item difficulty, item discrimination and distractor performance (incorrect answer choice). Problematic items are identified and sent back to appropriate test committees for evaluation and revision, if required.

The IRT item analysis is performed using the one-parameter (1-PL) logistic model. The 1-PL model describes the probability that candidates with a given ability level will respond correctly to an item as a function of item difficulty and their ability as measured by the exam in its entirety. Candidates with lower ability stand a lesser chance of answering the item correctly, while those with more ability are more likely to answer correctly. The mathematical expression for the 1-PL model is: (Hambleton, Swaminathan & Rogers, 1991):

$$P_i(\theta) = \frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}} \quad i = 1, 2, \dots, n$$

where

$P_i(\theta)$ is the probability that a randomly chosen candidate with ability θ answers item i correctly

b_i is the item i difficulty parameter

n is the number of items in the exam

e is a constant approximately equal to 2.718

The IRT analysis is performed using the Bilog-MG3 software (Zimowski et al, 1996). The statistic examined includes:

- *Item b-parameter estimate:* This estimate indicates the point on the IRT ability scale where the probability of a correct response is 0.5. The greater the value of the b -parameter estimate, the more difficult the item.

CTT analysis is performed using the Statistical Analysis System (SAS) and Bilog-MG3. The statistics examined include:

- *Item p-value*: This statistic indicates the proportion of candidates in the sample that answered the item correctly. The p -value ranges from 0.0 to 1.0. The higher the p -value, the easier the item.
- *Item-total correlation (point-biserial)*: This statistic is the correlation between the item score and the total test score and describes the relationship between performance on the specific item and performance on the total test. It indicates an item's discrimination power and its value ranges from -1.0 to +1.0. The higher the correlation, the better the item is at discriminating high-ability candidates from low-ability candidates. Items with negative correlations may point to serious problems with the item content (such as multiple correct answers or unusually complex content).
- The proportion of candidates choosing each answer option, including both the correct answer and incorrect answers (distractors) is also provided. It is desirable to have each answer option chosen by at least a few candidates.
- *Distractor-total correlation*: This statistic is the correlation between a distractor and the total test and describes the relationship between selecting an incorrect response for a specific item and performance on the entire test. A very low or negative value is desirable as more low ability candidates are expected to select these incorrect responses.

Each statistic provides some information about the characteristics of an item from an empirical perspective. These statistics are used to evaluate each item's psychometric quality and help detect any potential content-related issues. Items that fall into the following categories are not included in an item pool and are flagged for further review:

- p -value $< .05$ or p -value $> .95$
- Point-biserial $< .05$
- b -parameter < -5.5 , or b -parameter > 5.5

5.2 Item bank calibration

The MCCEE item bank was calibrated and scaled using the 1-PL IRT model described above. Prior to 2013, the items in the bank were calibrated using the item responses of all test takers gathered up to the time when the calibration was performed. In April 2013, following best practice, the item bank was recalibrated using only the item responses of first-time test takers between 2008 and 2012 (repeaters were excluded from the calibration sample). For the purpose of establishing a new scale for the bank, a concurrent calibration was implemented where b -parameters for all items (previously banked items and pilot items) were estimated simultaneously using the Bilog-MG3 software (Zimowski et al, 1996).

Concurrent calibration places item parameters on a common scale so that ability estimates from different administrations are comparable (Kang & Peterson, 2009; Kim, 2006; Kim, 2007). In 2016, the item bank was updated and recalibrated using candidate response data from January 2013 to May 2016. A fixed-parameter calibration (FPC) with simple transformation prior update (STPU) method (Kim, 2006) was used to link the scale of the new bank to the scale of the previous bank of items. Some items were excluded due to low discrimination power and/or because they were too easy or too difficult. The remaining items, along with their statistics, were uploaded to the bank.

Each year, pilot items need to be calibrated and scaled to the common bank scale once adequate data becomes available for these items. Due to the LOFT design, item exposure rates vary widely across items. To obtain an adequate sample size for the purposes of IRT calibration and scaling of pilot items, item responses from multiple administrations are combined excluding:

- Items with fewer than 100 responses as these may result in unstable parameter estimates
- Items with a p -value equal to zero (nobody answered the item correctly) or a p -value equal to one (everybody answered the item correctly) as parameters cannot be estimated (no variance)

A common-item, non-equivalent groups design is used, where all the operational items (i.e., counting items) are treated as anchor items to link the pilot items to the bank. Bilog-MG3 is first used to estimate b -parameter estimates for all items in separate calibrations. The new and banked b -parameter estimates for the anchor items are then used to estimate slope and intercept values using the IRT Mean-Mean (Kolen & Brennan, 2004) method to enable a linear transformation to put the b -parameter estimates of the pilot items on the scale of the bank. After scaling, pilot items that meet psychometric criteria are added to the item bank for future use.

For the purpose of LOFT test delivery as described in Section 5.3, items in the bank are classified into four difficulty levels based on their b -parameter estimates, with level 1 representing the easiest level and level 4 the most difficult level:

- Level 1: $b \leq -2.25$
- Level 2: $-2.25 < b \leq -0.75$
- Level 3: $-0.75 < b \leq 0.75$
- Level 4: $b > 0.75$

5.3 LOFT delivery

As indicated in Section 1, the MCCEE is administered using a computer-based, LOFT design and is delivered securely by Prometric, a test service provider. With the LOFT design, an exam form is constructed in real-time by selecting items from a large pool of operational

items each time a candidate takes the exam. Test security is enhanced because of the large number of forms the LOFT process can assemble. Each form is constrained by exam content specifications and psychometric criteria (test information target, item difficulty and item exposure parameters). Although each candidate receives a unique set of items, scores from all exams are comparable as all items in the pool are pre-calibrated and linked to a common scale established for the item bank from which the operational pool is drawn. The cut score is equivalent across exam forms.

When submitting items to Prometric to be used on an exam, the MCC provides the following information:

- The exam specifications including the upper and lower boundaries for each content category (minimum and maximum number of items allowed per content category), plus a weight (0-1.00) for each category. The weight value represents the proportion of the items from that category that should be included on the exam
- A large pool of items with IRT b -parameter estimates and their associated content categories
- The classification of item difficulty level ranging from one to four
- A list of enemy items (list of any two items that should not be included on the same exam form because their content overlaps or could provide answer clues to each other)
- A psychometric target for test information and standard errors of ability estimates (see Section 2.2, Figure 1)
- A list of pilot items assembled in six to seven packets of 30 items to be presented along with the operational items
- The length of the exam including the number of operational items and the number of non-scored pilot items

Based on this information, Prometric calculates an exposure control parameter for each operational item in the pool. The exposure control parameter represents the probability that an item will be selected for an exam. Items that best meet both the content specifications and the psychometric targets will have higher exposure control parameters than items that are less optimal at meeting these constraints. Items are selected for a candidate's exam from a large pool of items through randomization and optimization procedures. Items with higher exposure control parameters have a higher likelihood of being included in an exam form.

It is possible for some items to be exposed more often. This occurs especially in smaller categories in which certain items may have higher values toward meeting the psychometric target than other items in that category. To avoid overexposure of items, it is important that there be a sufficient number of items in all categories to prevent oversampling and overexposing some items.

The MCC monitors item exposure for each administration of the MCCEE and works closely

with Prometric to address any related issues. For a given administration, items are considered:

- “Overexposed” when seen by more than 50% of candidates
- “Underexposed” when seen by at least one candidate but less than 5% of candidates
- “Not exposed” when not seen by any candidates

5.4 Standard setting

Every few years, the MCC brings together a panel of Canadian physicians to define an acceptable level of performance and establish the pass score for the MCCEE through a standard-setting exercise. The panel then recommends its pass score to the C for approval.

In November 2016, the MCC conducted a rigorous standard-setting exercise with a diverse panel of 21 physicians from across the country. The method used is called the Bookmark Method, which has been widely used for multiple-choice question exams. Following the standard-setting exercise, the panel recommended a pass score of 261 on the current reporting scale of 50-500. This pass score was reviewed and approved by the EECC.

The new pass score of 261 was applied starting with the May 2017 session of the MCCEE and will remain in place until the next standard-setting exercise takes place.

Prior to May 2017, the pass score for the MCCEE was 250 on the reporting scale of 50-500. For candidates who took the MCCEE prior to May 2017, their final result (pass or fail) remains valid.

5.5 Scoring and score reporting

The 150 operational items that each candidate answers on the MCCEE are scored, but the 30 pilot items included in exam forms are not. The candidate ability θ is estimated using a 1-PL IRT model with a Bayes Expected A Posteriori (EAP) procedure. Thetas are scaled to have a mean of 0 and standard deviation of 1.0. The banked b -parameter estimates are used to estimate each candidate’s ability score θ along with their item response patterns.

Theoretically, the values of θ can range from $-\infty$ to $+\infty$, but practically, most of the θ values typically range from -3.0 to +3.0. To make it easier to communicate exam scores to candidates and other test users, the estimated θ score is linearly transformed onto a reporting scale to eliminate decimals and negative numbers. The reporting scale used for the MCCEE ranges from 50 to 500 with a standard deviation of 50. Transformed scores that are below 50 are adjusted to 50 and scores above 500 are adjusted to 500.

The θ cut score of -0.490 converts to a reported scale score of 261. Each candidate’s estimated θ score is converted to a reported score using the following equation:

$$\text{Reported score (rounded)} = 50 * (\theta + 0.490) + 261$$

In addition to providing candidates with their total score in the SOR, the MCC also provides supplemental graphical feedback via the SFR on the candidates' performance on the health group, clinician task and specialty area sub-categories to help them understand their strengths and weaknesses as assessed by the MCCEE. It is important to note that subscores have lower measurement precision than total scores as there are fewer items. The subscores are provided to individual candidates for feedback only and are not meant to be used by organizations for selection decision-making.

SECTION 6: EXAM RESULTS

Candidate performances for the five administrations in 2017 are summarized in this section. When applicable, historical data from previous years are included for reference purposes.

6.1 Candidate cohorts

In 2017, the MCCEE was administered in January, March, May, September and October/November to a total of 3,282 candidates in 198 cities in 61 countries. Table 6 summarizes the distribution of candidates per region and per cohort for the 2017 administrations.

Table 6: Distribution of candidates in 2017 by region

| Administration | REGION | | | | | | Total N |
|-------------------|--------------|-----------|------------|-----------|---------------|-----------|--------------|
| | Canada | | USA | | International | | |
| | N | % | N | % | N | % | |
| Jan. | 182 | 53 | 21 | 6 | 139 | 41 | 342 |
| March | 450 | 38 | 214 | 18 | 526 | 44 | 1,190 |
| May | 365 | 48 | 40 | 5 | 355 | 47 | 760 |
| Sept. | 200 | 52 | 13 | 3 | 171 | 45 | 384 |
| Oct. / Nov. | 274 | 45 | 37 | 6 | 295 | 49 | 606 |
| TOTAL 2017 | 1,471 | 45 | 325 | 10 | 1,486 | 45 | 3,282 |
| 2016 | 1,624 | 47 | 326 | 10 | 1,486 | 43 | 3,436 |
| 2015 | 1,770 | 46 | 356 | 9 | 1,690 | 44 | 3,816 |
| 2014 | 1,857 | 48 | 384 | 10 | 1,595 | 42 | 3,836 |
| 2013 | 1,835 | 50 | 422 | 12 | 1,412 | 38 | 3,669 |
| 2012 | 1,737 | 48 | 507 | 14 | 1,376 | 38 | 3,620 |

*Percentages do not total 100% due to rounding.

Table 7 presents the distribution of candidates who attempted the exam in various test centres in Canada in 2017.

Table 7: Distribution of candidates in Canadian test centres in 2017 by administration

| CENTRE | January | | March | | May | | September | | Oct./Nov. | | 2017 Total | |
|--------------------|----------|------------|------------|------------|------------|------------|--------------|----|-----------|----|------------|-----------|
| | N | % | N | % | N | % | N | % | N | % | N | % |
| Calgary | 20 | 11 | 30 | 7 | 47 | 13 | 16 | 8 | 16 | 6 | 129 | 9 |
| Edmonton | 18 | 10 | 38 | 8 | 30 | 8 | 22 | 11 | 33 | 12 | 141 | 10 |
| Halifax | 4 | 2 | 9 | 2 | 9 | 2 | 4 | 2 | 5 | 2 | 31 | 2 |
| Hamilton | 7 | 4 | 20 | 4 | 18 | 5 | 8 | 4 | 16 | 6 | 69 | 5 |
| London | 3 | 2 | 15 | 3 | 10 | 3 | 6 | 3 | 6 | 2 | 40 | 3 |
| Mississauga | 19 | 10 | 45 | 10 | 41 | 11 | 16 | 8 | 25 | 9 | 146 | 10 |
| Montreal | 17 | 9 | 39 | 9 | 49 | 13 | 28 | 14 | 38 | 14 | 171 | 12 |
| Ottawa | 7 | 4 | 18 | 4 | 15 | 4 | 13 | 7 | 20 | 7 | 73 | 5 |
| Regina | 0 | 0 | 3 | 1 | 1 | 0 | 3 | 2 | 5 | 2 | 12 | 1 |
| Saskatoon | 5 | 3 | 9 | 2 | 8 | 2 | 7 | 4 | 9 | 3 | 38 | 3 |
| St. John's | 1 | 1 | 4 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 7 | 0 |
| Toronto | 52 | 29 | 157 | 35 | 82 | 22 | 44 | 22 | 71 | 26 | 406 | 28 |
| Vancouver | 24 | 13 | 50 | 11 | 36 | 10 | 24 | 12 | 21 | 8 | 155 | 11 |
| Winnipeg | 5 | 3 | 13 | 3 | 17 | 5 | 9 | 5 | 9 | 3 | 53 | 4 |
| TOTAL | N | 182 | 450 | 365 | 200 | 274 | 1,471 | | | | | |
| | % | 12 | 31 | 25 | 14 | 18 | | | | | | |

*Percentages do not total 100% due to rounding.

Table 8 presents the distribution of candidates within major groups for each administration in 2017 as well as the total for the year.

Table 8: Distribution of candidates in 2017 by group and administration

| Candidate Group | Jan. | | March | | May | | Sept. | | Oct./Nov. | | 2017 Total | |
|-----------------------------|----------|------------|--------------|------------|------------|------------|--------------|----|-----------|----|--------------|-----------|
| | N | % | N | % | N | % | N | % | N | % | N | % |
| 1st-time test takers | 275 | 80 | 1,047 | 88 | 601 | 79 | 296 | 77 | 462 | 76 | 2,681 | 82 |
| Repeat test takers | 67 | 20 | 143 | 12 | 159 | 21 | 88 | 23 | 144 | 24 | 601 | 18 |
| English | 328 | 96 | 1,168 | 98 | 728 | 96 | 364 | 95 | 576 | 95 | 3,164 | 96 |
| French | 14 | 4 | 22 | 2 | 32 | 4 | 20 | 5 | 30 | 5 | 118 | 4 |
| TOTAL | N | 342 | 1,190 | 760 | 384 | 606 | 3,282 | | | | | |
| | % | 10 | 36 | 23 | 12 | 19 | | | | | | |

*Percentages do not total 100% due to rounding.

6.2 Overall Exam Results

Table 9 summarizes the descriptive statistics for the total score and pass rates for each cohort in 2017 as well as for the whole year. The scores are presented on the reporting scale ranging from 50 to 500, with a pass score of 250 (January and March) and a pass score of 261 (May, September and October/November).

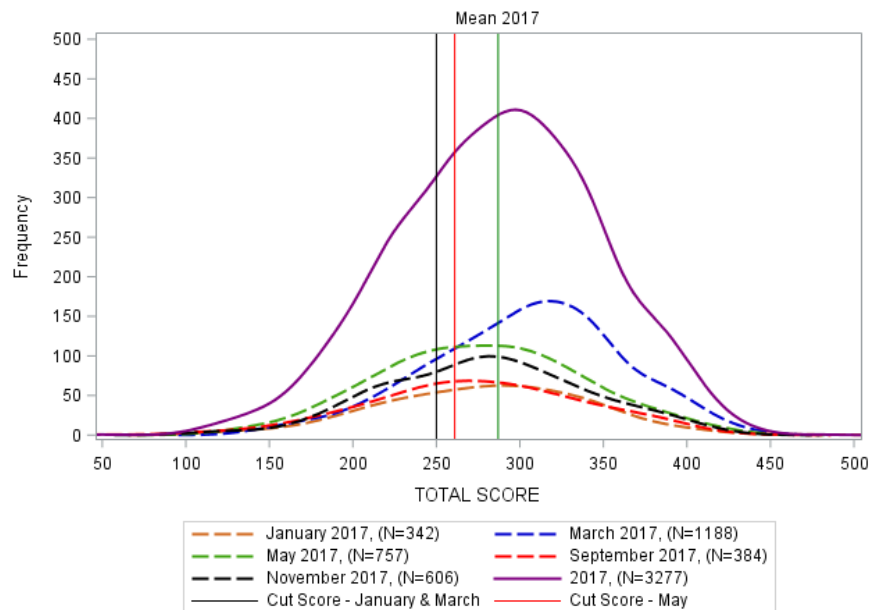
Table 9: Descriptive statistics for the total score and pass rates in 2017 by administration

| <i>Administration</i> | N | Min | Max | Mean | SD | PASS | |
|-----------------------|--------------|-----------|------------|------------|-----------|--------------|-----------|
| | | | | | | N | % |
| <i>January</i> | 342 | 119 | 478 | 279 | 60 | 236 | 69 |
| <i>March</i> | 1,188 | 132 | 443 | 303 | 57 | 970 | 82 |
| <i>May</i> | 757 | 116 | 437 | 277 | 60 | 447 | 59 |
| <i>September</i> | 384 | 50 | 435 | 274 | 64 | 222 | 58 |
| <i>Oct. / Nov.</i> | 606 | 105 | 427 | 279 | 60 | 382 | 63 |
| Total | 3,277 | 50 | 478 | 287 | 61 | 2,258 | 69 |

* Excluding candidates whose results were 'denied standing' or 'no standing'. The results 'denied standing' and 'no standing' are included in Tables 6, 7 and 8 as these tables did not report performance statistics.

Figure 2 displays the total score distributions on the reporting score scale for each cohort as well as for all candidates in 2017.

Figure 2: Total score distributions in 2017



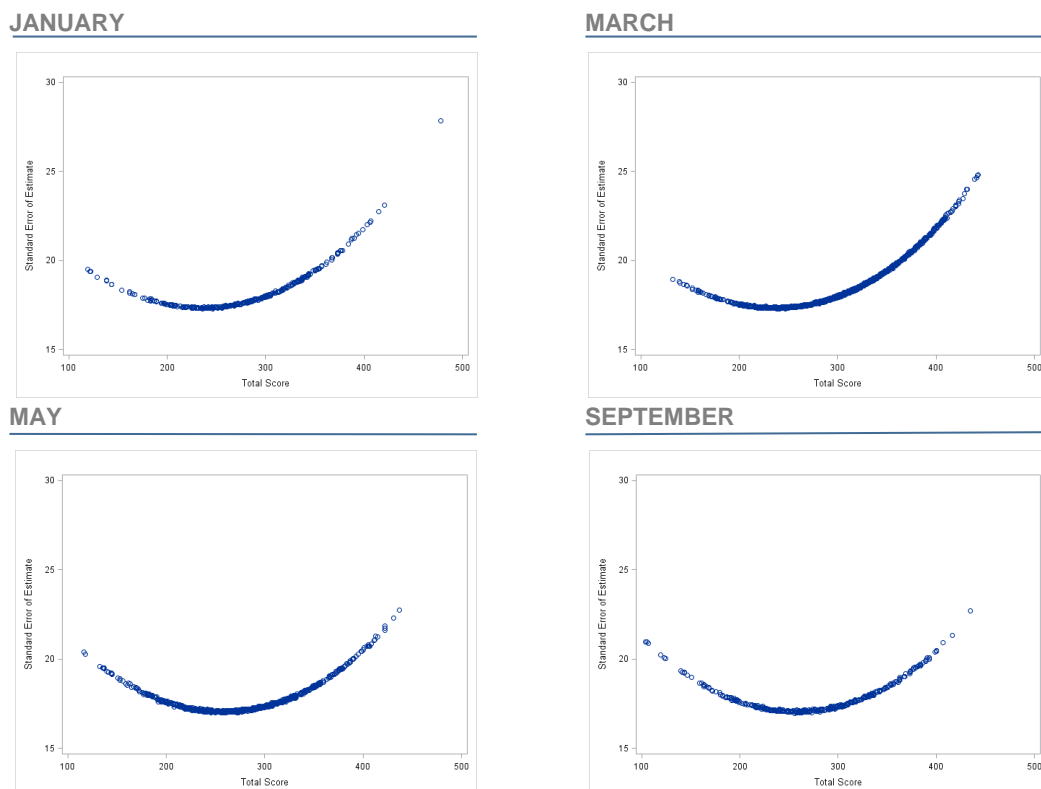
* Excluding candidates whose status was 'denied standing' or 'no standing'.

6.3 Reliability of exam scores and classification decisions

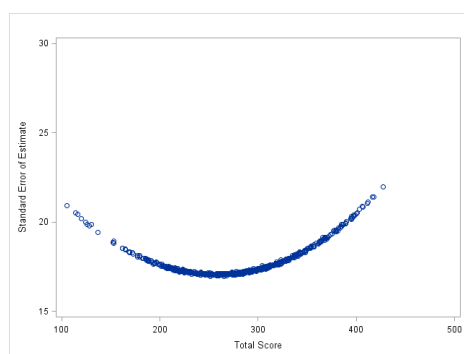
Test reliability refers to the extent to which the sample of items that comprises any exam accurately measures the intended construct. Reliability of the MCCEE can be assessed by examining the standard error of estimate (SEE) along the ability scale. The SEE indicates the precision with which ability is estimated at a given point on the ability scale and is inversely related to the amount of information provided by a test at that point (see Section 2.2 for an explanation of the test information function). The SEE values should be as small as possible so that measurement of the candidate's ability is as accurate as possible. In the IRT framework, the SEE serves the same purpose as the standard error of measurement (SEM) in CTT (Hambleton, Swaminathan & Rogers, 1991) except that the SEE varies with ability level in IRT whereas in CTT, one SEM is used to indicate overall measurement error.

Figure 3 displays the scatter plots of the SEE values along the ability scale (converted to the MCCEE reporting score scale) for the five cohorts in 2017. For each cohort, the plot shows that the ability estimates are less accurate towards 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 SEE is the lowest near the cut score, which indicates the highest precision of ability estimates, thus supporting more accurate and consistent pass or fail decisions.

Figure 3: Distributions of standard errors of the ability estimates for 2017 cohorts



OCTOBER / NOVEMBER



* Excluding candidates whose status was 'denied standing' or 'no standing'.

A critical concern for a high-stakes exam such as the MCCEE is the pass or fail decision. Reliability of the MCCEE can also be assessed by examining the consistency and accuracy of pass or fail decisions based on exam scores. Decision consistency and decision accuracy can be estimated using the Livingston and Lewis (1995) procedure, which is used in many high-stakes testing programs. Decision consistency is an estimate of the agreement between the pass or fail classifications on potential parallel forms of the exam. Decision accuracy is an estimate of the agreement between the pass or fail classifications based on observed exam scores and those that would be based on their true score (expected average score if the candidate could be tested an infinite number of times).

Table 10 shows the decision consistency and decision accuracy estimates along with the associated false positive and false negative rates. The estimated false positive rate indicates the expected proportion of candidates who pass based on their observed score but who should fail based on their true ability. The estimated false negative rate indicates the expected proportion of candidates who fail based on their observed score but who should pass based on their true ability. As indicated in Table 10, both the decision consistency and the decision accuracy estimates for the five 2017 administrations are very high; false positive and false negative rates are within an acceptable range.

Table 10: Estimates of decision consistency and decision accuracy in 2017

| | January | March | May | September | Oct./Nov. |
|-----------------------------|-------------|-------------|-------------|-------------|-------------|
| Decision consistency | 0.87 | 0.90 | 0.85 | 0.85 | 0.86 |
| False positive | 0.04 | 0.03 | 0.05 | 0.04 | 0.05 |
| False negative | 0.05 | 0.04 | 0.06 | 0.06 | 0.06 |
| Decision accuracy | 0.91 | 0.93 | 0.89 | 0.90 | 0.90 |
| False positive | 0.04 | 0.03 | 0.05 | 0.05 | 0.05 |
| False negative | 0.05 | 0.04 | 0.06 | 0.06 | 0.05 |

* Excluding candidates whose status was 'denied standing' or 'no standing'.

6.4 Candidate performance by sub-category

In Table 11, descriptive statistics are presented for total exam scores as well as for subscores based on three different but inter-related classification systems: health groups, clinician tasks and specialty areas for the 2017 candidates. Each domain within each classification system is sampled a number of times, with some being measured by a large number of questions and others by a smaller number of questions. Note that the questions overlap across the three classification systems.

Table 11: Descriptive statistics for the total score and subscores in 2017

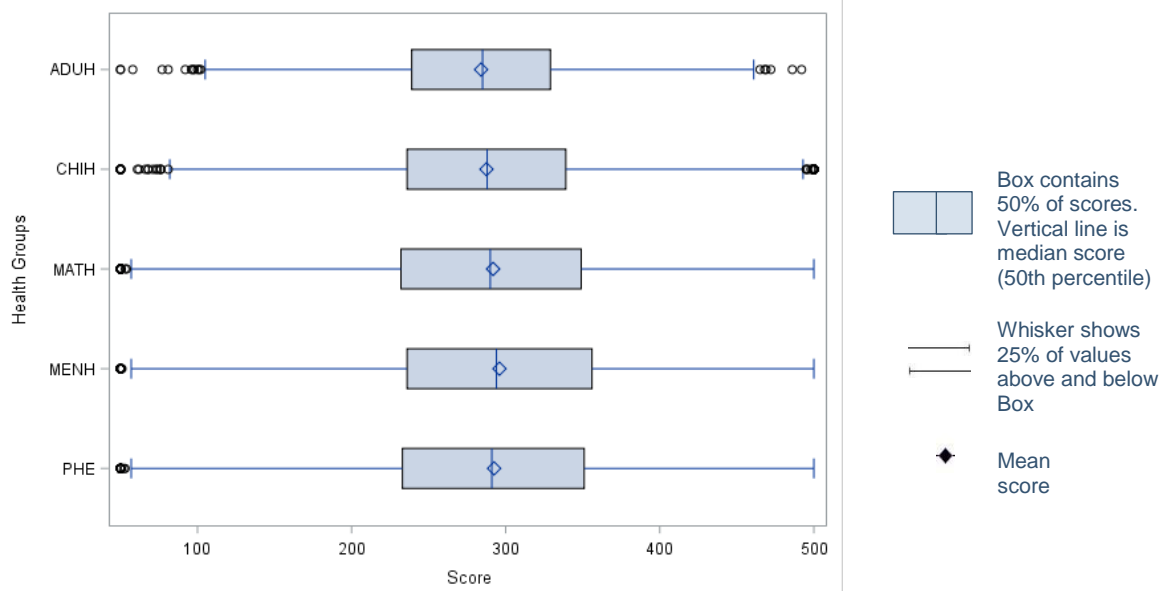
| | | Min | Max | Mean | SD |
|------------------------|-----------------------------------|-----------|------------|------------|-----------|
| TOTAL SCORE | | 50 | 478 | 287 | 61 |
| Health groups | Adult Health | 50 | 492 | 284 | 65 |
| | Child Health | 50 | 500 | 288 | 81 |
| | Maternal Health | 50 | 500 | 292 | 91 |
| | Mental Health | 50 | 500 | 296 | 87 |
| | Population Health and Ethics | 50 | 500 | 292 | 91 |
| Clinician tasks | Data gathering | 50 | 500 | 285 | 75 |
| | Data interpretation and synthesis | 50 | 500 | 286 | 70 |
| | Management | 50 | 490 | 289 | 62 |
| Specialty areas | Medicine | 50 | 500 | 285 | 82 |
| | Obstetrics & Gynecology | 50 | 500 | 290 | 74 |
| | Surgery | 50 | 500 | 283 | 71 |
| | Pediatrics | 50 | 500 | 288 | 81 |
| | Psychiatry | 50 | 500 | 296 | 87 |
| | Population Health and Ethics | 50 | 500 | 292 | 91 |

* Adult Health includes Medicine, Surgery and Obstetrics & Gynecology

* Excluding candidates whose status was 'denied standing' or 'no standing'.

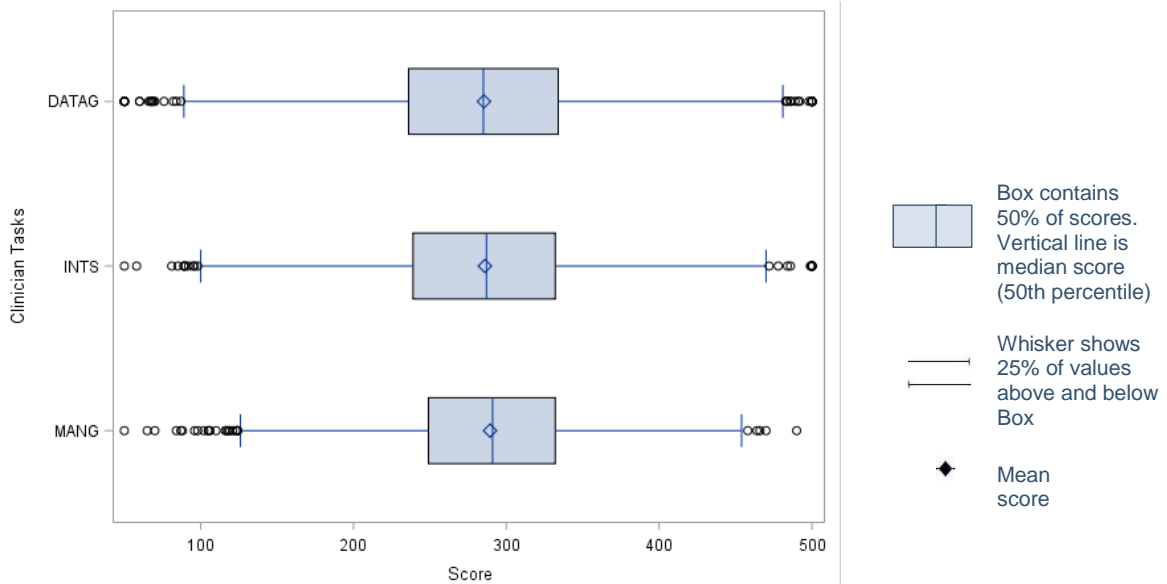
Figures 4 through 6 show subscore distributions and the profiles of candidate performances in the health group, clinician task and specialty area domains respectively for 2017. The box for each domain indicates the range for the middle 50% of candidate scores. The vertical line represents the median or 50th percentile score for that domain. Each line to the right or left of the box represents the remaining 25% of the domain score above or below the middle 50%. The mean domain score is shown by the diamond. Overlap between the boxes indicates that candidate performances in those domains did not differ significantly.

Figure 4: Subscore distributions for health groups in 2017



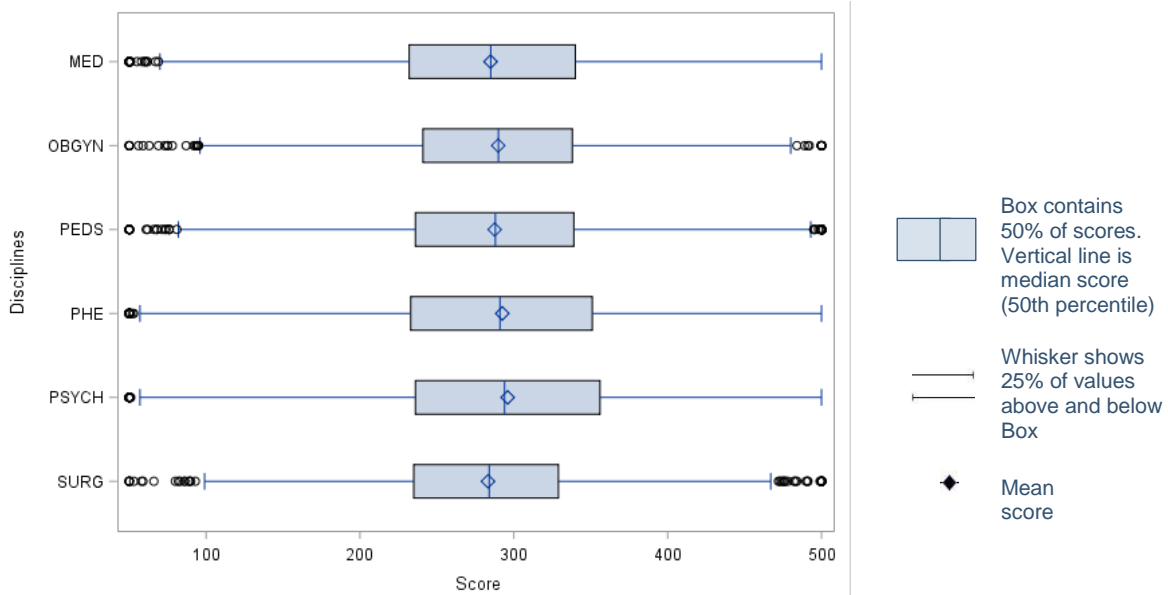
* ADUH: Adult Health, CHH: Child Health, MATH: Maternal Health, MENH: Mental Health, PHE: Population Health and Ethics
 * Excluding candidates whose status was 'denied standing' or 'no standing'.

Figure 5: Subscore distributions for clinician tasks in 2017



* DATAG: Data gathering, INTS: Data interpretation and synthesis, MANG: Management
 * Excluding candidates whose status was 'denied standing' or 'no standing'.

Figure 6: Subscore distributions for specialty areas in 2017



* MED: Medicine, PEDS: Pediatrics, PHE: Population Health and Ethics, PSYCH: Psychiatry, OBGYN: Obstetrics and Gynecology, SURG: Surgery
 * Excluding candidates whose status was 'denied standing' or 'no standing'.

6.5 Exam results by candidate group

Table 12 presents descriptive statistics and pass rates for each candidate group in 2017: first-time test takers, repeat test takers, and candidates who took the exam in English or French.

Table 12: Descriptive statistics and pass rates in 2017 by candidate group

| GROUP | N | | Min | Max | Mean | SD | PASS | |
|-------------------------------|---------------|----|-----------|------------|------------|-----------|--------------|-----------|
| | N | % | | | | | N | % |
| All candidates | 3,277* | | 50 | 478 | 287 | 61 | 2,258 | 69 |
| <i>First-time test takers</i> | 2,678 | 82 | 50 | 478 | 297 | 60 | 2,055 | 77 |
| <i>Repeat test takers</i> | 599 | 18 | 105 | 357 | 240 | 41 | 203 | 34 |
| <i>English</i> | 3,162 | 96 | 50 | 478 | 288 | 61 | 2,212 | 70 |
| <i>French</i> | 115 | 4 | 104 | 379 | 244 | 51 | 46 | 40 |

* Excluding candidates whose status was 'denied standing' or 'no standing'.
 * Percentages do not total 100% due to rounding.

6.6 Comparison of prior exam performance

Table 13 presents pass rates of each cohort in 2017 and those of the previous years.

A pass score of 250 on the reporting scale ($\theta = -0.704$ on the IRT ability scale) was applied prior to May 2017. As of May 2017, a pass score of 261 on the reporting scale ($\theta = -0.490$) was applied. It should be noted that in the summer of 2016, the item bank was re-calibrated using first-time takers only. Starting in May 2017, the new IRT parameters have been applied to the scoring of MCCEE candidate performance following the recalibration of the items in the MCCEE item bank.

Table 13: Pass rates of each 2017 administration and the previous five years

| Year | Administration | N | Overall Pass Rate (%) | First-Time Taker Pass Rate (%) |
|-------------|----------------|---------------|-----------------------|--------------------------------|
| 2017 | January | 342 | 69 | 78 |
| | March | 1,188 | 82 | 86 |
| | May | 757 | 59 | 68 |
| | September | 384 | 58 | 68 |
| | November | 606 | 63 | 72 |
| | TOTAL | 3,277 | 69 | 77 |
| 2016 | January | 393 | 70 | 76 |
| | March | 1,243 | 82 | 85 |
| | May | 826 | 66 | 73 |
| | September | 409 | 67 | 73 |
| | November | 562 | 66 | 73 |
| | TOTAL | 3,433* | 72 | 78 |
| 2015 | January | 436 | 66 | 72 |
| | March | 1,259 | 80 | 85 |
| | May | 992 | 63 | 70 |
| | September | 493 | 71 | 80 |
| | November | 631 | 63 | 69 |
| | TOTAL | 3,811 | 70 | 77 |
| 2014 | January | 379 | 64 | 70 |
| | March | 1,168 | 78 | 83 |
| | May | 1,072 | 67 | 74 |
| | September | 529 | 68 | 74 |
| | November | 689 | 65 | 72 |
| | TOTAL | 3,837 | 70 | 76 |
| 2013 | January | 435 | 77 | 86 |
| | March | 513 | 79 | 83 |
| | May | 982 | 80 | 85 |
| | September | 1,035 | 90 | 92 |
| | November | 705 | 63 | 70 |
| | TOTAL | 3,670 | 79 | 84 |

* Excluding candidates whose status was 'denied standing' or 'no standing'.

6.7 Item exposure analysis

As mentioned in Section 5.3, the items in each exam form for each candidate are selected based on item exposure control parameters that reflect how well an item meets test specifications and the psychometric target. As items in an exam form are delivered through randomization and optimization procedures, some items may be more highly exposed than others. The MCC monitors each administration for item exposure and addresses the issue together with Prometric.

Table 14 presents the items for the five administrations in 2017.

Table 14: Item exposure in 2017

| Administration | Overexposed | Underexposed | Unexposed | Number of Candidates |
|--------------------|-------------|--------------|-----------|----------------------|
| January (Pool 7) | 0 | 115 | 181 | 342 |
| March (Pool 7) | 0 | 197 | 101 | 1,190 |
| May (Pool 8) | 0 | 152 | 146 | 760 |
| September (Pool 8) | 0 | 104 | 193 | 384 |
| November (Pool 8) | 0 | 120 | 176 | 606 |

6.8 Candidate survey

Every year, a survey is administered to candidates at the end of the exam seeking feedback on their test-taking experience. The survey is used for quality improvement purposes. Table 15 presents a total of 2132 candidates who answered the survey in 2017. Please note that survey data for the September administration was not captured due to a technical data issue.

Table 15: Candidate Survey Results (2017)

Q 1. How satisfied are you with the staff's helpfulness at this centre?

| A – Very Satisfied | B – Satisfied | C – Dissatisfied | NR ¹ |
|--------------------|---------------|------------------|-----------------|
| 72% | 27% | 1% | 755 |

Q 2. How satisfied were you with the performance of the testing system during your examination?

| A – Very Satisfied | B – Satisfied | C – Dissatisfied | NR ¹ |
|--------------------|---------------|------------------|-----------------|
| 51% | 44% | 5% | 761 |

Q 3. How satisfied are you with the total experience of taking your examination at this Prometric testing centre?

| A – Very Satisfied | B – Satisfied | C – Dissatisfied | NR ¹ |
|--------------------|---------------|------------------|-----------------|
| 52% | 45% | 3% | 754 |

Q 4. Overall, how would you rate the format of the examination (including such factors as screen layout, and ease of use)?

| A – Very Satisfied | B – Satisfied | C – Dissatisfied | NR ¹ |
|--------------------|---------------|------------------|-----------------|
| 40% | 54% | 6% | 766 |

Q 5. How would you rate the time allotted to complete the examination?

| A – Far too little | B – Too little time | C – About the correct amount of time | D – Time to spare | E – Much time to spare | NR ¹ |
|--------------------|---------------------|--------------------------------------|-------------------|------------------------|-----------------|
| 4% | 24% | 62% | 8% | 2% | 766 |

Q 6. How would you rate the quality of the images presented with the questions?

| A – Very Satisfied | B – Satisfied | C – Dissatisfied | NR ¹ |
|--------------------|---------------|------------------|-----------------|
| 39% | 50% | 11% | 765 |

Q 7. How would you rate the clarity of the instructions you were provided on completing this examination?

| A – Very Satisfied | B – Satisfied | C – Dissatisfied | NR ¹ |
|--------------------|---------------|------------------|-----------------|
| 53% | 45% | 2% | 773 |

Q 8. How would you rate this examination as an appropriate test of your medical knowledge?

| A – Very Satisfied | B – Satisfied | C – Dissatisfied | NR ¹ |
|--------------------|---------------|------------------|-----------------|
| 19% | 62% | 18% | 786 |

¹ NR refers to the number of surveys without a response to that question. The percentage represents the calculated average of the four sessions with rounding, therefore, some may not total 100.

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.
- Cizek, G. J. (ed.) (2001). *Setting Performance Standards: Concepts, Methods and Perspectives*. New Jersey: Lawrence Erlbaum Associates Inc.
- Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications Inc.
- Kang, T. & Peterson, N. S. (2009). *Linking Item Parameters to a Base Scale*. ACT Research Report Series 2009-2.
- Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43, 355-381.
- Kim, J. (2007). A comparison of calibration methods and proficiency estimators for creating IRT vertical scales. PhD (Doctor of Philosophy) thesis, University of Iowa, 2007.
- Kolen, M. J., & Brennan, R. L., (2004). *Test equating, scaling, and linking: methods and practice*. (2nd ed.) New York, NY: Springer.
- Linn, R. L., (2000). The standards for educational and psychological testing: Guidance in test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (1st ed., pp. 27-28). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement*, 32, 179-197.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3–19.
- Messick, S. (1989). Validity. In *Educational Measurement* (3rd ed., p. 610). Macmillan USA.
- Tong, Y., & Michael, J. K. (2010). Scaling: An ITEMS Module. *Educational Measurement: Issues and Practice*, 29 (4), 39–48.
- Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (1996). BILOG-MG3. SSI Inc.

APPENDIX A: LIST OF COUNTRIES WHERE THE MCCEE IS OFFERED

NORTH AMERICA

| Country | # of Centres |
|---------------|--------------|
| Canada | 13 |
| United States | 336 |
| Mexico | 4 |

EUROPE

| Country | # of Centres |
|----------------|--------------|
| Armenia | 1 |
| Austria | 1 |
| Bulgaria | 1 |
| Croatia | 1 |
| Czech Republic | 1 |
| Finland | 1 |
| France | 4 |
| Georgia | 1 |
| Germany | 5 |
| Greece | 2 |
| Hungary | 1 |
| Ireland | 1 |
| Italy | 3 |
| Kazakhstan | 1 |
| Latvia | 1 |
| Lithuania | 1 |
| Luxembourg | 1 |
| Netherlands | 1 |
| Poland | 1 |
| Portugal | 1 |
| Romania | 1 |
| Russia | 2 |
| Slovenia | 1 |
| Spain | 2 |
| Switzerland | 1 |
| Turkey | 4 |
| Ukraine | 1 |
| United Kingdom | 13 |
| Uzbekistan | 1 |

SOUTH AMERICA

| Country | # of Centres |
|--------------------|--------------|
| Argentina | 2 |
| Bolivia | 1 |
| Brazil | 7 |
| Chile | 1 |
| Colombia | 2 |
| Dominican Republic | 1 |
| Guatemala | 1 |
| Peru | 1 |
| Venezuela | 1 |

ASIA PACIFIC

| Country | # of Centres |
|-------------|--------------|
| Australia | 2 |
| Bangladesh | 1 |
| China | 17 |
| Hong Kong | 2 |
| Indonesia | 2 |
| India | 19 |
| Japan | 9 |
| Korea | 7 |
| Malaysia | 1 |
| Nepal | 1 |
| Pakistan | 3 |
| Philippines | 3 |
| Singapore | 1 |
| Taiwan | 3 |
| Thailand | 1 |

AFRICA

| Country | # of Centres |
|--------------|--------------|
| Botswana | 1 |
| Ghana | 1 |
| Kenya | 1 |
| Mauritius | 1 |
| Nigeria | (closed) |
| South Africa | 2 |
| Tanzania | 1 |
| Uganda | 1 |
| Zimbabwe | 1 |

MIDDLE EAST

| Country | # of Centers |
|----------------------|--------------|
| Egypt | 2 |
| Israel | 2 |
| Jordan | 1 |
| Kuwait | 1 |
| Lebanon | 2 |
| Saudi Arabia | 3 |
| United Arab Emirates | 1 |
| West Bank | 1 |
| | |
| | |
| | |
| | |
| | |
| | |
| | |

APPENDIX B: STATEMENT OF RESULTS (SOR) EXAMPLE



MEDICAL COUNCIL
OF CANADA LE CONSEIL MÉDICAL
DU CANADA

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

**MEDICAL COUNCIL OF CANADA
EVALUATING EXAMINATION
STATEMENT OF RESULTS**

November 6, 2017


| | | | |
|----------------------------|----------------------|--------------------------------|------|
| Candidate name: | XXXXXXX, XXXXX XXXXX | Your final result: | Pass |
| MCC candidate code: | XXXXXXXXXX | Your total score: | 320 |
| Examination date: | 2017-09-08 | Score required to pass: | 261 |


On behalf of the Evaluating Examination Composite Committee, I am writing to inform you of your final result on the Medical Council of Canada Evaluating Examination (MCCEE) that took place on the above-mentioned date.

Your total score, which represents your overall performance, is reported as a scaled score ranging from 50 to 500. Your final result (e.g., pass/fail) is based on your total score relative to the score required to pass. Additional information, including the mean and standard deviation, is available from the [MCCEE Scoring web page](#).

Supplemental feedback on your examination performance is reported to you in a separate document within your [physiciansapply.ca](#) account.


Please accept my best wishes for future success.


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



mcc.ca
physiciansapply.ca
inscriptionmed.ca

APPENDIX C: SUPPLEMENTAL FEEDBACK REPORT (SFR) EXAMPLE



MEDICAL COUNCIL OF CANADA LE CONSEIL MÉDICAL DU CANADA

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

SUPPLEMENTAL FEEDBACK REPORT

| | | | |
|----------------------------|-------------------|--------------------------|------------|
| Candidate name: | Xxxxxx, Xxxx Xxxx | Examination: | MCCEE |
| MCC candidate code: | XXXXXXXXXX | Examination date: | 2017-09-08 |
| Your total score: | 320 | | |

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

Figures 1 to 3 display your performance for domains based on three different but interrelated classification systems: first by Health Group (i.e., Adult Health, Child Health, Maternal Health, Mental Health, Population Health and Ethics); then by Clinician Task (i.e., Data Gathering, Data Interpretation and Synthesis, Management); and finally by Discipline (i.e., Medicine, Obstetrics and Gynecology, Pediatrics, Population Health and Ethics, Psychiatry, Surgery). Each domain within each classification system is sampled a number of times, with some being measured by a large number of questions and others by a smaller number of questions. Note that the questions overlap across the three classification systems.

To help you better understand your performance, your subscore for each domain is shown along with the mean score of candidates who were first-time takers of the MCCEE since May 2017 and who passed. The standard error of measurement (SEM) associated with each of your subscores represents the expected variation in your subscore if you were to take this examination again with a different set of questions covering the same or similar domains. Small differences in subscores or overlap between SEMs are indicative that performance in those domains was relatively similar. Likewise, overlap between the SEM for a domain subscore and the mean score of first-time takers who passed, within a given domain, signifies that performance is similar to the mean.

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

For more information, please visit the [MCCEE Scoring web page](#).

Report date: 2017-11-06 1/3

mcc.ca
physiciansapply.ca
inscriptionmed.ca

SUPPLEMENTAL FEEDBACK REPORT

Figure 1. MCCEE Score Profile: Your performance by Health Group

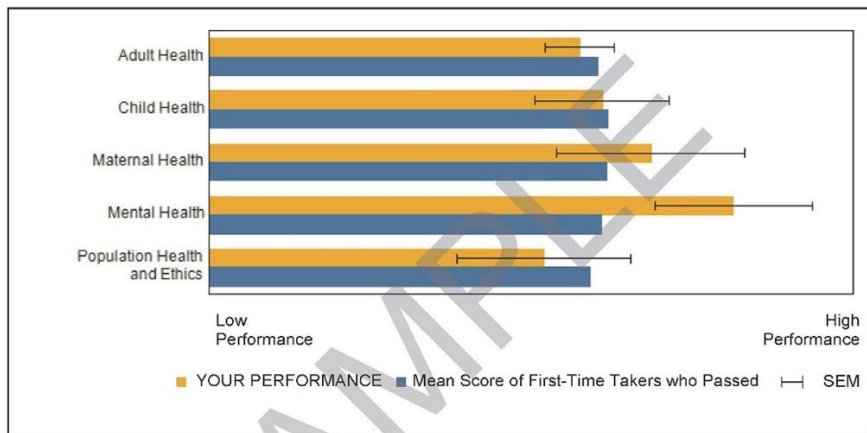
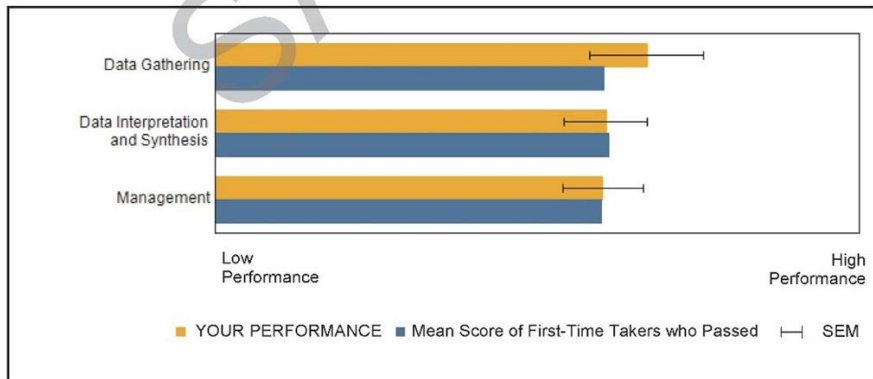


Figure 2. MCCEE Score Profile: Your performance by Clinician Task

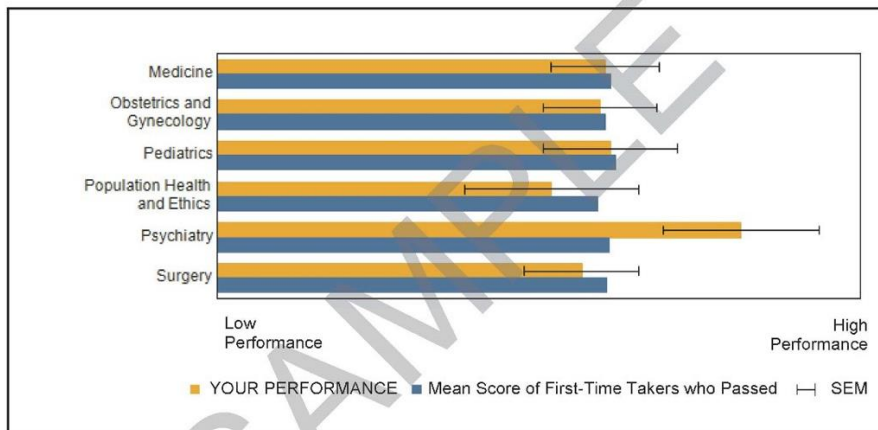


Report date: 2017-11-06 2/3 Xxxxxx, Xxxx Xxxx / XXXXXXXX

mcc.ca
 physiciansapply.ca
 inscriptionmed.ca

SUPPLEMENTAL FEEDBACK REPORT

Figure 3. MCCEE Score Profile: Your performance by Discipline



Report date: 2017-11-06

3/3

Xxxxxx, Xxxx Xxxx / XXXXXXXX

mcc.ca
physiciansapply.ca
inscriptionmed.ca

