MEDICAL COUNCIL
OF CANADA

LE CONSEIL MÉDICAL
DU CANADA

Technical report on the standard-setting exercise for the Medical Council of Canada Qualifying Examination Part I

Psychometrics and Assessment Services

September 2018

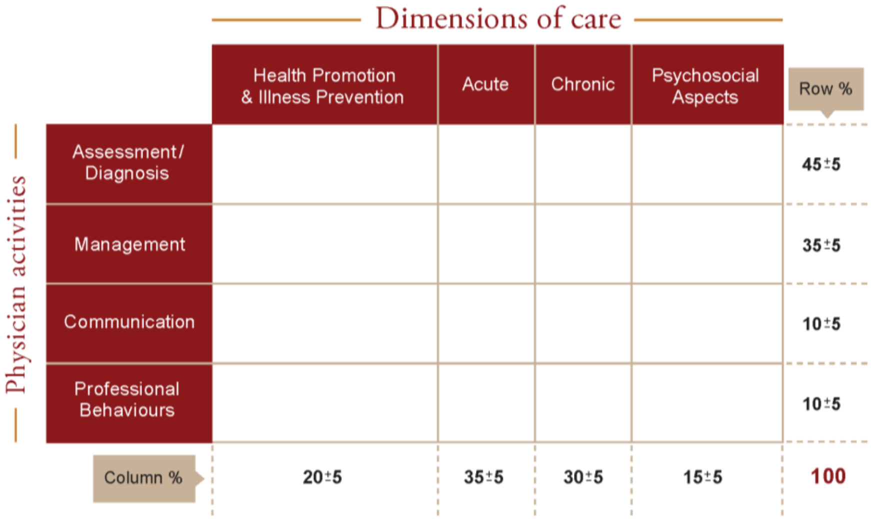
Table of Contents

1. Background and purpose

Standard setting is a critical component of any high-stakes assessment program, particularly for licensing and certification decisions in the health professions. We need to assure the public that licence and certificate holders possess the required knowledge, skills and attitudes necessary for safe and effective patient care. Standard setting is a process used to define an acceptable level of performance in the competency domains targeted by an examination. The resulting conceptual standard is operationalized as a numerical pass score (also known as a cut score) that is used to make classification decisions (e.g., pass/fail, grant/withhold a credential, award/deny a licence).

The Medical Council of Canada Qualifying Examination (MCCQE) Part I is a computer-delivered examination which assesses basic medical knowledge and skills expected to be mastered at the end of medical school. It is composed of a four-hour Multiple-Choice Questions (MCQ) component and a three and a half hour Clinical Decision-Making (CDM) component. In spring 2018 the Medical Council of Canada (MCC) launched a new Blueprint for the MCCQE Part I to reflect the changing requirements for physicians to practice safe and effective patient care. The new exam Blueprint assesses two broad categories: Dimensions of Care and Physician Activities. Table 1 displays the MCCQE Part I exam specifications. The MCQ component consists of 210 questions and the CDM component consists of 38 cases with a combination of short-menu and short-answer write in questions.

Table 1. MCCQE Part I exam specifications



Dimensions of care

Physician activities

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

The previous pass score for the MCCQE Part I was established on the previous Blueprint in 2014. It is best practice to review the standard and the pass score every three to five years or sooner if there is a change to the exam, such as a new blueprint. This is to ensure that the standard is appropriate and reflects the current level of competency to practise in the profession; thereby protecting public interest and keeping current with the evolvement of the exam and test-taker population, and advancements in medicine and medical education.

On June 18-19, 2018, a panel of 22 physicians from across Canada met at the MCC's offices in Ottawa to participate in a standard-setting exercise for the MCCQE Part I. Staff from the Psychometrics and Assessment Services (PAS) directorate facilitated the meeting, with support from the Chief Medical Education Officer and staff in Evaluation Bureau (EB). The purpose of the meeting was to arrive at a recommended pass score for subsequent consideration and approval by MCC's Central Examination Committee (CEC), a body that is responsible for overseeing the MCCQE Part I including the development and maintenance of exam content and the approval of exam results.

In this report, we summarize the process, procedures and results of the two-day exercise that led to the recommendation and approval of a new pass score for the MCCQE Part I.

2. Procedures

2.1. SELECTING A STANDARD-SETTING METHOD

Several standard-setting methods are appropriate for MCQ exams (Cizek & Bunch, 2007). We selected the Bookmark method as our primary method based on several considerations.

- First, the MCCQE Part I is a criterion-referenced exam for which a pass score should be defined as an acceptable amount of knowledge that candidates must possess, or an acceptable level of performance they need to demonstrate given the intended use of the exam. A pass or fail status should be determined by comparing an individual candidate's performance to a performance standard regardless of the performance of other candidates. Therefore, a criterion-referenced method of standard setting such as the Bookmark method is most appropriate for the MCCQE Part I.
- Secondly, the MCCQE Part I is composed of MCQs and CDM cases that include both short-menu and short-answer, write-in items. The Bookmark method is appropriate for setting standards using both types of items, selected response and short-answer, write-in items. This method has been widely used for setting standards on licensure and certification examinations. The Bookmark method is a test-centered, criterion-referenced method where expert judges review test items and provide judgments as to the adequate level of performance on those items. Conversely, for examinee-centered approaches for performance exams, other methods for setting standards are preferred, for example, Borderline Group or Contrasting Group methods (Kane, 1998).
- Thirdly, the Bookmark method is a convenient way to connect a cut score to the Rasch model, which is used to calibrate items and assemble test forms for the MCCQE Part I. The Rasch model characterizes examinee ability and item difficulty simultaneously, making it possible to order items by the ability needed to have a specific probability of

success and to map the items on the ability scale. In this way, candidates with scores near the location of specific items can be inferred to possess the abilities required to respond successfully to those items with the specified probability.

- Fourthly, the Bookmark method simplifies the cognitive complexity required of standard-setting judges and is relatively easy to use compared to other methods.
- Finally, we have used the Bookmark method successfully for setting a standard on the MCCQE Part I in the past and on the Medical Council of Canada Evaluating Examination (MCCEE), which is similar to the MCCQE Part I.

We also chose to complement the Bookmark method with the Hofstee method. The Bookmark and Hofstee methods are described below.

2.1.1. Bookmark method

The Bookmark method is an item mapping procedure where items are ordered from easiest to most difficult and panelists are asked to place a bookmark at the point at which they believe a minimally proficient candidate would no longer correctly answer subsequent items presented. De facto, this corresponds to the cut score for each panelist.

Specifically, the Bookmark method is a procedure where items are presented one per page and ordered from easiest to most difficult based on operational data. Standard-setting panelists review each item in the order presented and consider the likelihood of a correct response by a minimally competent candidate (see section 2.4.1.2). For each item, each panelist makes a judgement on whether a minimally competent candidate would have a good chance of answering the item correctly. For our purpose, we defined “good chance” as having at least 0.67 response probability (RP67) or 2/3 odds of answering the item correctly. In completing these judgments, panelists consider multiple factors, including: (1) the knowledge being assessed by that item, (2) the difficulty level of that item and, (3) the definition of the minimally competent candidate. Panelists make this judgment for every item until they reach a point in the exam booklet where they feel a minimally competent candidate would no longer have a 67 per cent chance of answering items correctly. They then place a “bookmark” on that page; hence why it is called the Bookmark method. A more detailed description of the Bookmark method is provided in Cizek & Bunch (2007). It is important to note that candidates may be able to answer some items correctly beyond that page or bookmark. Even by random guessing, with five answer choices, they have a 20 per cent chance of answering an item correctly. However, the panelists are instructed to place their bookmark where the minimally competent candidate’s chance would fall below 0.67 probability or 2/3 odds.

Each individual panelist’s cut score corresponds to the Rasch ability level (i.e., RP67) associated with the bookmarked item. A cut score is derived from the average or median of the cut scores across panelists and then by panels. This process can be repeated in two or three rounds. Impact data (i.e., pass/fail rate of the cut score) are usually presented for discussion after each round to help panelists to understand the consequences of their recommendation.

2.1.2. Hofstee method

The use of criterion-referenced approaches sometimes may lead to unacceptable outcomes in the absence of political considerations associated with the decision (De Champlain, 2013). To ensure the standard set by using the Bookmark method is ‘in touch with reality’, we also used the Hofstee method to check its reasonableness from a policy perspective. The Hofstee method is a “compromise” method that uses both a holistic judgment on an acceptable cut score (criterion-referenced) and an acceptable failure rate (norm-referenced), concurrently. It derives a cut score based on answers to the following four questions that panelists are asked to address based on their expertise and experience in the field, knowledge of the test content and objective of the examination, as well as their understanding of the test-taker population:

- What is the highest percent correct cut score that would be acceptable, even if every candidate attains that score?
- What is the lowest percent cut score that would be acceptable, even if no candidate attained that score?
- What is the maximum failure rate that would be acceptable?
- What is the minimum failure rate that would be acceptable?

Panelists’ answers to the first two questions provide absolute information for a criterion-referenced standard based on exam content whereas answers to the last two questions provide relative information to define a norm-referenced standard based on candidates’ performance. The answers to each question are averaged across panelists and then plotted in a graph displaying the cumulative percentage of candidates who would fail at each point along the score scale (see section 3.3). The Hofstee method is usually not used as a standalone method. For our purpose, we used it to complement the Bookmark method and provide a “reality check” on the standard set using the Bookmark method. A more detailed description of the Hofstee method is provided in Cizek & Bunch (2007) and Hofstee (1983).

2.2. SELECTING PARTICIPANTS AND ASSIGNING INTO PANELS

Selecting a panel of well-qualified panelists is an important step to ensure the validity of a standard-setting process and the resulting cut score. Due to the inherent subjectivity of any standard-setting process, best practice dictates the selection of a panel that broadly represents the target examination population, with respect to background and educational characteristics (De Champlain, 2013).

In July 2017, the MCC sent out an email invitation to many individuals and groups from across the country to solicit interest in participating in our standard-setting exercise. This solicitation resulted in over 250 interested physicians, each of whom completed a demographic information form. The original invitation email and demographic form are included in Appendix A.

Based on the demographic information provided, the MCC selected 22 participants and assigned them to two subpanels that were matched as closely as possible on key demographic variables, including: (1) number of years in practice (post-residency), (2) geographic region, (3) gender, (4) medical specialty, and (5) ethnic background. The main purpose of using two subpanels was to assess the replicability of the cut score across two parallels of independent groups of physicians. This provides evidence of the replicability and generalization (i.e., validity) of the recommended cut score. In addition, smaller subpanels may foster more discussions as they allow each participant more opportunity to share his or her perspective. Table 2 summarizes the demographic composition of the two subpanels.

Table 2: Demographic information by standard-setting subpanel

Variable of interest	Group	Subpanel 1	Subpanel 2	Total
Number of years in practice post-residency	0-10	4	4	8
	11-30	5	6	11
	30+	2	1	3
Geographic region	Western Canada	3	3	6
	Ontario	5	5	10
	Quebec	2	2	4
	Maritimes	1	1	2
Gender	Male	7	5	12
	Female	4	6	10
Specialty	Primary care*	5	7	12
	Other care	6	4	10
Ethnic background	Caucasian	7	7	14
	Visible minority	4	4	8

* Primary care is defined as the category of healthcare professionals who provide direct first-contact services such as family physicians and pediatricians (canada.ca/en/health-canada/services/primary-health-care/about-primary-health-care.html).

2.3. PREPARING MATERIALS FOR THE STANDARD-SETTING EXERCISE

2.3.1. *Test form selected for Ordered Item Booklet*

There were multiple test forms used in the Spring 2018 MCCQE Part I administration. One form was selected to be used for the Ordered Item Booklet (OIB) that was judged to be typical in content and met strict psychometric specifications. This test form had 210 active items that were used for the standard-setting exercise. The 210 items were assembled in a 222-page OIB with one item per page ordered from the easiest to the most difficult based on item difficulty parameter estimates.

2.3.2. *RP67s*

With the Bookmark method, panelists make a judgement on whether a minimally competent candidate has a good chance of answering each item correctly. As indicated in section 2.1.1, we defined “good chance” as having at least 0.67 response probability (i.e., RP67) or 2/3 odds. Though we considered other probability levels (e.g., RP50), we decided to use RP67 as this is typically used by other testing programs. RP67 is consistent with the mastery notion for a criterion-referenced exam (i.e., you need to have an ability that will give you greater than 0.50 probability of answering an item correctly to be considered as having mastered the content knowledge assessed by the item). In addition, it is a relatively easy value for standard-setting judges to understand (especially when expressed as 2/3 odds). The ability level needed to have a 0.67 response probability of answering an item correctly was calculated using the formula (Cizek & Bunch, 2007):

$$\theta_i=\beta_j+.708$$

where θ_i and β_j represent examinee ability and item difficulty, respectively. The RP67 values were calculated for each of the 210 items selected for the standard-setting exercise.

2.3.3. *Item map*

An item map was prepared that included information about the item order (i.e., page number of each item) in the OIB, item ID, answer key, RP67 value and content classification for each item.

2.3.4. *Practice Booklet and Practice OIB*

Panelists were provided with a Practice test and a practice OIB to familiarize themselves with the MCCQE Part I and prepare themselves for the standard-setting activity. The Practice test consisted of 53 items (43 MCQ items and 6 CDM cases with 10 questions). The items were randomly ordered in a booklet with one to two items per page and the items represented a range of difficulty levels and content domains.

The Practice OIB consisted of 54 pages – 52 pages for 52 dichotomous items and two pages for a polytomous item scored as 0, 1, 2. Items were ordered from the easiest to the most difficult. For each item on each page of the Practice OIB, we also provided its item ID, answer key and RP67 value.

2.3.5. Background materials

Panelists were provided with the agenda and two research papers prior to the standard-setting exercise (De Champlain, 2013; Karantonis & Sireci, 2006). The research papers provided panelists with an overview of standard setting and the Bookmark method. Panelists were asked to read the papers in advance to gain a preliminary understanding of standard setting and the Bookmark method.

2.4. ACTIVITIES DURING THE TWO-DAY SESSION

The agenda for the two-day meeting is provided in Appendix B. Day 1 was devoted to training the panelists whereas Day 2 was devoted to the actual standard-setting exercise.

2.4.1. Day 1 – Training and practice

Day 1 was devoted exclusively to the training of the panelists as the credibility and defensibility of the new cut score relies heavily on extensive training of the panelists. We began the meeting with a welcome and a round-table introduction of facilitators and panelists, as well as an overview of the purpose of the meeting. We told panelists specifically that their task was to recommend a pass score, not to make a final decision, and that we would submit their recommendation to the CEC for consideration and approval. We then provided an overview of the MCCQE Part I including its purpose, content, format, scoring, score reporting, psychometric model, exam delivery model and intended test-taker population. We followed this by an overview of the standard-setting exercise including its purpose, process, selection and training of panelists, criterion- and norm-referenced frameworks and common methodologies. We also provided a brief explanation of the Bookmark method.

2.4.1.1. Familiarizing judges with the MCCQE Part I

To familiarize the judges with the type of questions and difficulty level of the MCCQE Part I, we gave them an hour to review the Practice Booklet (see section 2.3.4) and answer the 53 sample questions that were presented in random order. We then provided the panelists with the answer key to self-score their answers without sharing their score with other judges. Afterwards, the panelists discussed their perceived difficulty level of the questions and the range of content coverage keeping in mind the purpose of the MCCQE Part I and its target test-taker population.

2.4.1.2. Defining the minimally competent candidate

A critical step in any standard-setting exercise is to define the target candidate for the proficiency level targeted by the examination. For the MCCQE Part I, the target is the minimally competent candidate entering supervised practice in Canada. MCC's Chief Medical Education Advisor and the CEC approved the definition of the minimally competent candidate provided in Appendix C. In the afternoon of Day 1, panelists reviewed the definition and engaged in a discussion lead by MCC's Chief Medical Education Advisor describing the qualities of this individual and how they are distinguishable from the incompetent candidate's. Panelists were asked to share their thoughts, envision some minimally competent candidates, discuss their characteristics, capabilities, things they may have difficulty doing and what distinguishes the "minimally competent" from the "incompetent" candidates. The intention was to help panelists converge on a unified conceptualization of the minimally competent candidate given the purpose of the MCCQE Part I so that they could recommend a meaningful cut score. The discussion continued until everyone was satisfied that they understood who the minimally competent candidate was. We asked panelists to keep in mind the definition and the image of the minimally competent candidate consistently throughout the two-day exercise.

2.4.1.3. Practice using the Bookmark and Hofstee methods

After panelists familiarized themselves with MCCQE Part I content and reached a common understanding of the definition of the minimally competent candidate, we provided a step-by-step training on how to use the Bookmark method to set a cut score and reviewed the four Hofstee questions as described in section 2.1.2. We divided panelists into two pre-assigned subpanels and assigned each panel to a different room. We provided the panelists with an opportunity to practise setting a cut score using the Bookmark method using the 53 items in the Practice OIB. All items were the same as in the Practice Booklet that panelists completed in the morning with the exception that the items in the Practice OIB were ordered by difficulty from easiest to most difficult. Each panelist's task was to individually review each item in the order presented and provide a judgement on whether a minimally competent candidate would have at least a 0.67 probability or 2/3 odds of answering the item correctly. We asked each panelist to place their bookmark on the page beyond which a minimally competent candidate would have a probability of less than 0.67 or 2/3 odds of correctly answering all items. We asked them to record their bookmark page on a Bookmark Form (see Appendix D) as well as provide responses to the four Hofstee questions (see Appendix E). During the practice, panelists were also provided an item map for the Practice OIB which included information about the item order, item ID, answer key, RP67 value, and item content classification.

The cut score is determined by the median practice bookmark cut scores from the individual judges within each subpanel and then the average of the two subpanels (see section 2.1.1). We presented these practice results to the full panel for discussion, questions and clarifications to ensure that judges would have a good sense of the process of recommending a cut score and the impact of a cut score on pass/fail rates.

By the end of Day 1, panelists developed a very good understanding of the purpose, content and difficulty level of the MCCQE Part I, the definition of the minimally competent candidate, the standard-setting process and the Bookmark and Hofstee methods.

2.4.2. Day 2 – Standard-setting exercise

Day 2 started with a brief recap of the previous day's activities where we reminded panelists of the key points about the Bookmark method. We then proceeded to conducting two rounds of the standard-setting exercise.

2.4.2.1. Round 1 (preliminary round)

For Round 1, we split panelists into two subpanels and placed them in two different rooms. A psychometrician facilitated each subpanel. We provided panelists with an OIB containing 210 items for standard setting, ordered by difficulty from easiest to most difficult (see section 2.3.1). We also provided an item map to panelists that included information about the item order, item ID, answer key, RP67, and item content classification. We instructed panelists to review the items in the OIB in the order presented, starting from page 1 and to place a bookmark at the point beyond which they felt a minimally competent candidate would no longer have a 0.67 probability or 2/3 odds to correctly answer all items. Panelists were given three hours to independently provide a bookmark judgment and record their bookmark page on a Bookmark Form (see Appendix D). During this activity, each panelist had a printed copy of the definition of the minimally competent candidate as well as the purpose of the MCCQE Part I to allow panelists to refer to the definition and purpose of the exam at all times while making judgement on each item.

After panelists completed their bookmark judgments, we asked them to answer the four Hofstee questions as described in section 2.1.2 (see Appendix E). Specifically, we asked panelists to specify the highest and lowest percent scores as well as the highest and lowest failure rates they believed would be reasonable for the MCCQE Part I based on their holistic judgment and knowledge of the purpose of the exam, and definition of the minimally competent candidate and their experience.

We collected the completed Bookmark and Hofstee forms and during the panelists' lunch break, PAS staff tallied the bookmarks and calculated median cut scores by individual panelist, subpanel and full panel as well as Hofstee results for each subpanel and the full panel. We also calculated the impact of the full panel's cut score on failure rate using Canadian medical graduates first-time test takers on the MCCQE Part I from Spring 2018.

We then presented the results and impact data from Round 1 to all panelists before splitting them once again into two groups in different rooms. They were given 15 minutes for discussion within each subpanel. Based on the impact data, some panelists felt they were too lenient in terms of what they expected the minimally competent candidate would be able to master while others felt they were too harsh. We then brought the two subpanels back together in one room with each subpanel appointing a spokesperson who brought a summary of their discussion to the full group. The full panel discussed and shared further thoughts on the process and outcomes. For comparison, panelists were also shown historical failure rates of first-time test takers and all test takers in 2016 and 2017, based on the previous cut score that was established in 2014.

The Round 1 exercise provided judges with an opportunity for realistic practice in full scale. Round 1 results, impact data and discussions helped to calibrate the panelists towards a better understanding of the process, a more unified idea about the cut score and potential consequences of their judgment. It also became clear to panelists why they needed to have a common understanding of the definition of the minimally competent candidate and to keep it in mind while making judgments on each item. With the information learned and skill developed from Round 1, panelists were better prepared for Round 2, the final round.

2.4.2.2. Round 2 (final round)

In Round 2, we split panelists into the same subpanels and assigned each panel to a separate room. Within each subpanel, they repeated the same exercise as in Round 1. That is, they independently provided Bookmark and Hofstee judgments using the OIB for standard setting and recorded their bookmark pages and answers to the four Hofstee questions using the forms provided. However, we told them that they did not need to start from page 1 (i.e., the easiest item); instead, they could narrow their focus on items they were previously unsure of or items near their Round 1 bookmark page (e.g., 15 items before and 15 items after). Panelists were told they can keep the same bookmark as Round 1 or change it. Again, we reminded them to keep the definition of the minimally competent candidate in mind while making judgments on the likelihood of answering an item correctly. They were given one and a half hours to complete this activity.

At the end of this activity, we collected the completed Bookmark and Hofstee forms. MCC staff calculated cut scores for individual panelists and subpanels and the full panel and created graphs and tables to show the impact of their cut scores on failure rates using the performance data of Canadian first-time test takers from the spring 2018 administration.

We then presented the results and impact data from Round 2 to the full panel. For comparison purposes, we again showed the historical failure rates of first-time test takers and all test takers in 2016 and 2017. We provided panelists with the opportunity to briefly discuss the Round 2 cut score, the standard-setting process used to derive it and any potential impacts on future MCCQE Part I candidates.

2.4.2.3. Calculation of the final recommended Bookmark cut score

An individual panelist’s cut score corresponded to the RP67 value for the item on their bookmarked page (i.e., the ability required to have a 0.67 probability of a correct response as expressed on an IRT θ scale). The subpanel’s cut score is the median RP67 value. The reason for using median instead of mean was that it is less affected by extreme values or outliers. Finally, the two cut scores from the two subpanels were averaged to obtain the full panel’s cut score.

3. Results

3.1. BOOKMARK RESULTS

Table 3 presents a summary of the Bookmark cut scores for each subpanel and the full panel. Overall there was more variability in the Bookmark cut scores for Subpanel 1 relative to Subpanel 2; however, this difference was larger in Round 1 than Round 2. For each subpanel and each round, we computed the Standard Error of Judgment (SEJ) which is an estimate of the variability that we would expect if the same judging process was repeated by many different panels of similar composition. We then constructed 95 per cent confidence intervals around their cut scores using SEJ for each subpanel. In the final round, the 95 per cent confidence intervals for Subpanel 1 (0.54, 0.95) and Subpanel 2 (0.46, 0.77) overlapped suggesting they did not differ significantly.

The θ cut score of 0.682 derived from Round 2 became the standard-setting panel’s final recommended cut score. A θ score of 0.682 translates to a scale score of 226 on the MCCQE Part I reporting scale of 100-400.

Table 3: Bookmark cut scores

		Cut score (*θ*)	Min	Max	SD	SEJ
Round 1	Subpanel 1	0.87	-0.24	1.39	0.48	0.36
	Subpanel 2	0.44	-0.69	0.78	0.46	0.35
	Full panel	0.658				
Round 2	Subpanel 1	0.75	0.44	1.32	0.27	0.21
	Subpanel 2	0.62	0.05	0.78	0.21	0.16
	Full panel	0.682				
Final cut score		**0.682**				

3.2. IMPACT DATA

As indicated earlier, we computed the impact of cut scores on failure rates using performance data from 2,810 Canadian Medical Graduate first-time test takers in Spring 2018. These results are presented in Table 4. The final round (Round 2) had a slightly higher failure rate than Round 1 but the difference was less than one percent. For comparison, historical failure rates are also shown for 2016-2017. As shown, the new cut score led to a similar failure rate as previous years.

Table 4: Failure rates by round and candidate cohort

	Recommended cut score		CMG first-time test takers	All test takers
	θ	Reported		
Round 1 (preliminary)	0.658	224	4.0%	19.1%
Round 2 (final)	0.682	226[1]	4.7%	20.2%
Historical failure rates				
Spring 2017	-0.223	427[2]	4.6%	21.9%
Spring 2016			3.2%	20.6%

3.3. HOFSTEE RESULTS

Table 5 summarizes the Hofstee results computed by averaging panelists' answers to the four Hofstee questions within each subpanel and for the full panel. The results from the two rounds are very similar.

Table 5: Summary of Hofstee results by round and subpanel

	Statistic	Subpanel 1	Subpanel 2	Full panel
Round 1	Min. acceptable Percentage cut score	55	50	**53**
	Max. acceptable Percentage cut score	71	69	**70**
	Min. acceptable Failure rate	2	2	**2**
	Max. acceptable Failure rate	16	8	**12**
Round 2	Min. acceptable Percentage cut score	55	50	**52**
	Max. acceptable Percentage cut score	68	68	**68**
	Min. acceptable Failure rate	2	2	**2**
	Max. acceptable Failure rate	16	8	**12**

[1] Scale of 100 to 400 with a mean of 250 and a standard deviation of 30

[2] Scale of 50 to 950 with a mean of 500 and a standard deviation of 100

Figure 1 displays the average answers from the full panel in Round 2 (as reported in Table 4) plotted against a cumulative percentage of candidates who would fail at each point along the θ ability scale using the performance data of Canadian first-time test takers from the Spring 2018 MCCQE Part I administration. Panelists felt that the cut score should not be lower than 0.14 and not higher than 1.05. Similarly, they indicated that the failure rate should be at least 2 per cent but not higher than 12 per cent. The coordinates (max. cut score and min. failure rate) and (min. cut score and max. failure rate) are linked by a green, dotted line. The point of intersection between this line and the cumulative frequency distribution corresponds to a cut score of close to 0.72 if we drew a vertical line down from this point to the horizontal axis. The cut score from the Bookmark method was slightly lower at 0.682. The Hofstee method would result in a higher failure rate than the Bookmark method (5.7 per cent vs 4.7 per cent failure rate for 2018 Canadian first-time test takers respectively). As indicated earlier, the Hofstee method was not our primary method for setting the standard for the MCCQE Part I; it was used for a "reality check" of the standard set by using the Bookmark method. The results indicate the Bookmark cut score was consistent with panelists' global judgment of the cut score and failure rate.

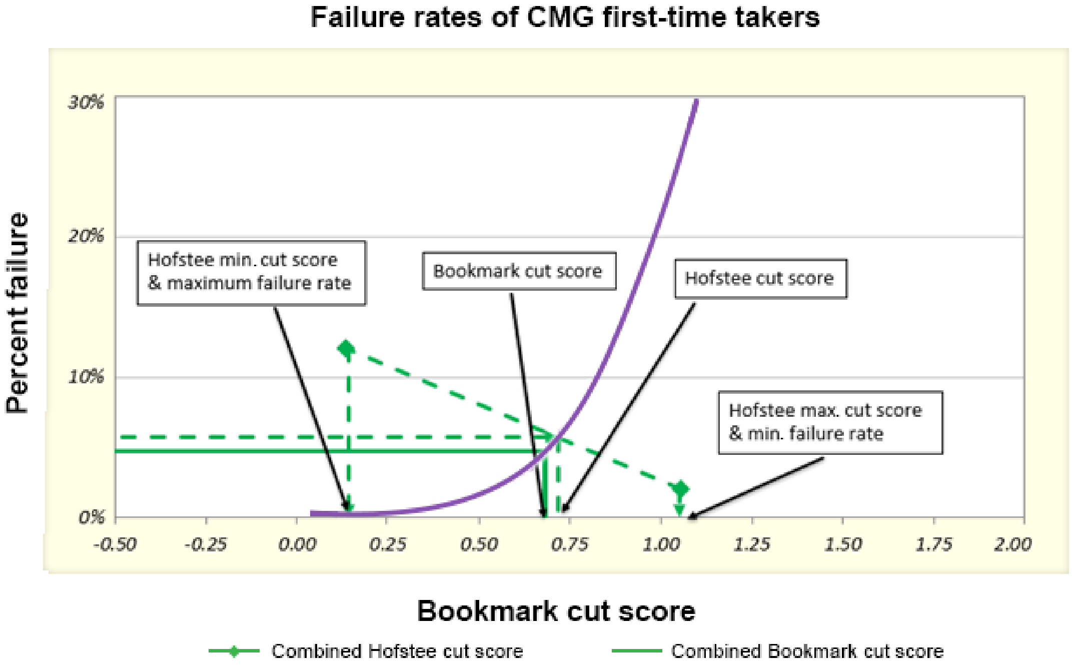


Figure 1: Average results from the full panel in Round 2

3.4. FINAL RECOMMENDED PASS SCORE

After a rigorous two-day standard-setting exercise, the panel of 22 physicians recommended a new pass score of 0.682 on the θ scale (226 on reporting scale) that was subsequently brought forward to the CEC for consideration and approval.

3.5. POST-SESSION SURVEY

Panelists were provided with a survey at the end of the standard-setting exercise, so they could provide anonymous feedback to MCC staff. Full results of the survey are presented in Appendix F. All 22 panelists responded to the survey. In summary, the survey results indicated the following:

- At the beginning of the session, we provided an overview of the standard-setting process. We surveyed panelists about the clarity of the information provided. All respondents found the overall evaluation of the training for setting a pass to be excellent (77.3 per cent), very good (13.6 per cent), good (4.6 per cent) or fair (4.6 per cent).
- Central to the standard-setting exercise is the notion of the minimally competent candidate. All panelists felt they were very clear (72.7 per cent) or clear (27.3 per cent) on the definition of the minimally competent candidate for the MCCQE Part I.
- We devoted a significant amount of time and effort to training panelists on the Bookmark procedure to ensure a common understanding of what was expected of them before they engaged in the actual exercise. In Round 1, respondents found the information regarding the Bookmark method to be very good (63.6.8 per cent), good (31.8 per cent) or poor (4.6 per cent). Panelists had a better understanding of the Bookmark method in Round 2 as they found the information to be very good (81.8 per cent), good (13.6 per cent), or fair (4.6 per cent).
- When asked about the length of time provided for completing the activity, most of the panelists said the provided time was about right (77.3 per cent) or too much time was provided (22.7 per cent).
- At the end of each round, we presented impact data to show the consequences of their preliminary cut score on failure rates. Respondents found impact data and subsequent discussions to be very helpful (77.3 per cent), helpful (18.2 per cent) or not helpful at all (4.6 per cent) in facilitating the panel to arrive at a defensible pass score.
- Finally, and most importantly, panelists indicated they were very confident (68.2 per cent), confident (27.3 per cent), or somewhat confident (4.6 per cent) in the final recommended cut score. None of the respondents indicated a lack of confidence.

4. Conclusions

Several findings highlight our confidence in the standard-setting process and the resulting pass score.

1. The recommended cut scores from Subpanel 1 and Subpanel 2 converged in Round 2. This indicates that the training provided to panelists in Round 1 and Round 2 further reinforced their understanding of the standard-setting process.

2. The 95 per cent intervals around the cut score constructed using the standard error of judgment for Subpanel 1 (0.54, 0.95) and Subpanel 2 (0.46, 0.77) indicate very similar ranges and significant overlap between the two subpanels. This provides evidence to support the replicability of the cut score if different panels were used to follow the same process.

3. The recommended Bookmark cut score was within the acceptable range defined by the Hofstee method based on judges' holistic judgment. This indicates that the criterion-referenced cut score derived using the Bookmark method is realistic and consistent with holistic judgments.

4. The results of the post-session survey indicate confidence in the training provided.

5. Panelists expressed high confidence in the standard-setting process and the final recommended pass score as indicated by the post-session survey results.

These findings provide evidence to support the reliability and validity of the standard-setting process and that the resulting recommended pass score is defensible from both psychometric and holistic perspectives.

The recommended pass score was presented to the CEC on June 25, 2018, along with an overview of the standard-setting process, and the impact data. The CEC unanimously approved the recommended θ pass score of 0.682 (226 on reporting scale) for the MCCQE Part I. The new pass score was implemented for Spring 2018.

References

Cizek, G. J. & Bunch, M. B. (2007). Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests (55-189). Thousand Oaks, CA: Sage.

De Champlain, A. F. (2013). Standard setting methods in medical education. In T. Swanwick (Ed.). Understanding Medical Education: Evidence, Theory and Practice. (305-316). Chichester, West Sussex: John Wiley & Sons, Ltd.

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

Karantonis, A. & Sireci, S. G. (2006). The Bookmark Standard Setting Method: A Literature Review. Educational Measurement: Issues and Practice, 25(1), 4-12.

Kane, M. (1998). Choosing Between Examinee-Centered and Test-Centered Standard- Setting Methods, Educational Assessment, 5 (3), 129-145.

APPENDIX A:
Demographic information sheet

The information requested below is being collected to help the MCC obtain a pan-Canadian representative panel to recommend a passing score on the MCC Part I Examination. This information will only be used to select the panel members so that we can represent the diversity of physicians across the country. The information will not be linked in any way to the collection of data for setting the passing score. A reminder that the meeting will take place on June 18 and 19, 2018 therefore we are asking panelists to be available on those two days.

Please provide your name and contact information and check a box next to each of the questions. The form can be sent by mail or electronically.

Name: ____

Email: ____________________________________ Phone number: _________________

Mailing address: ____

____

____

1. Number of years in practice post residency:

1-5 years	□
6-10 years	□
11-20 years	□
21-30 years	□
More than 30 years	□

1. Number of years' experience supervising residents:

1-5 years	□
6-10 years	□
11-20 years	□
21-30 years	□
More than 30 years	□

2. Do you have experience supervising Canadian Medical Graduates?

Yes □

No □

3. Have you ever been a member of a Medical Council test committee?

Yes □

No □

4. Country of medical training (post graduate training):

Canada □

Other □

5. Region of the country in which you live:

Alberta □

British Columbia □

Manitoba □

Maritimes □

Ontario □

Quebec □

Saskatchewan □

Territories □

6. First Language:

English □

French □

Other (__________________________) □

7. Gender:

Male □

Female □

8. Ethnicity:

Asian □

Black □

Caucasian □

First Nations □

Hispanic □

9. Medical Specialty:

Pediatrics □

Internal medicine □

Psychiatry □

Obstetrics and gynecology □

Surgery □

Family medicine □

Other □

10. Type of community in which you work:

Urban □

Rural □

11. Type of care setting:

Hospital-based □

Community-based □

APPENDIX B: Agenda

Medical Council of Canada
Qualifying Examination Part I standard-setting exercise
June 18-19, 2018

MCC Office: 1021 Thomas Spratt Place
Maude Abbott/Thomas Roddick meeting rooms (first floor)

DAY 1: June 18, 2018

TIME	ACTIVITIES	LEAD
08:00	**Breakfast (and registration)**	
08:30	Welcome and introductions	
08:40	Security video and business code of conduct	
08:50	Review agenda and objectives	
09:00	Overview of standard setting	
09:45	Overview of MCCQE Part I	
10:00	**Break**	
10:15	Panelists review and answer Practice test (subpanels in two rooms)	
11:15	Panelists self-score Practice test and discussion	
11:45	**Lunch**	
12:30	Develop common understanding of the definition of the minimally competent candidate entering residency	
13:30	Training of Bookmark method	
14:00	Round 0: Split into two subpanels and two rooms • Panelists independently practise Bookmark method using Ordered Item Booklet (OIB) for Practice test and provide Hofstee judgments	
15:00	**Break**	
15:00	Data entry and calculation	
15:15	Post-bookmark training discussion and clarification	
15:30	Round 0 feedback (Practice test): Bring subpanels into one room • Present Round 0 results and impact data	
16:00	Wrap-up of Day 1 / Overview of Day 2	
16:30	**End of Day 1**	
18:00	**Dinner**	

DAY 2: June 19, 2018

TIME	ACTIVITIES	LEAD
08:00	**Breakfast**	
08:30	Round 1: Split into two subpanels and two rooms • Panelists independently provide Bookmark and Hofstee judgments using the OIB for standard setting	
11:30	**Lunch**	
11:30	Data entry and calculation	
12:30	Round 1 Feedback: Bring subpanels into one room • Present Round 1 results and impact data	
13:00	Subpanels in two rooms and discuss impact data	
13:15	Subpanels in one room for whole panel discussion	
13:30	Round 2: Subpanels in two rooms • Panelists independently provide Bookmark and Hofstee judgments using the OIB for standard setting	
15:00	**Break**	
15:00	Data entry and calculation	
15:30	Round 2 Feedback: Bring subpanels into one room • Present Round 2 results and impact data	
16:00	Complete post-standard-setting exercise survey	
16:15	Wrap-up	
16:30	**End of Day 2**	

APPENDIX C:
Defining borderline performance and the minimally competent candidate

The “minimally competent” candidate entering supervised practice has just enough knowledge and skills to provide safe and effective patient care, no more, no less. A “minimally competent” candidate’s performance is acceptable, despite gaps in their knowledge and clinical decision-making skills.

APPENDIX D:
Form to document a bookmark for each round

Panel: ____

Panelist: ____

Standard setting for the MCCQE Part I
The Bookmark method

Please indicate the page number of the item on which you placed your bookmark. It is the item for which, in your judgment, a minimally competent candidate would have 0.67 probability (a 2/3 chance) of correctly answering all the items up to that point and their chance would go below 0.67 beyond that point.

Please initial after each round:

Round	Bookmark Page	Initials
1		
2		

APPENDIX E:
Form to document a Hofstee for each round

Panel: ____

Panelist: ____

Standard setting for the MCCQE Part I
The Hofstee method

Please answer the following 4 questions, once after each round:

1. What is the highest percent correct cut score that would be acceptable, even if every candidate attains that score?

   Practice: ______ Round 1: ______ Round 2: ______

2. What is the lowest percent correct cut score that would be acceptable, even if no candidate attains that score?

   Practice: ______ Round 1: ______ Round 2: ______

3. What is the maximum failure rate that would be acceptable?

   Practice: ______ Round 1: ______ Round 2: ______

4. What is the minimum failure rate that would be acceptable?

   Practice: ______ Round 1: ______ Round 2: ______

APPENDIX F:
Part I standard setting 2018 – post-session survey summary

1. Which panel did you participate in? (Select ONE)

Response	Percentage	Count
Panel 1	50.0%	11
Panel 2	50.0%	11
Total responses	**100.0%**	**22**

2. How clear did you find the information regarding the overview of the MCCQE Part I that was provided on the morning of Day 1?

Response	Percentage	Count
Very clear	63.6%	14
Clear	27.3%	6
Somewhat clear	9.1%	2
Not clear	0%	0
Total responses	**100.0%**	**22**

3. During training on Monday morning, did you feel the discussion of the "minimally competent" candidate on the MCCQE Part I was helpful?

Response	Percentage	Count
Yes, very helpful	72.7%	16
Yes, helpful	27.3%	6
Yes, somewhat helpful	0%	0
Not at all helpful	0%	0
Total responses	**100.0%**	**22**

4. Following discussion on Monday morning, how clear were you about the description of the "minimally competent" candidate on the MCCQE Part I as you began the task of practising to set a passing score on Day 1?

Response	Percentage	Count
Very clear	54.6%	12
Clear	27.3%	6
Somewhat clear	18.2%	4
Not clear	0%	0
Total responses	**100.0%**	**22**

5. How would you judge the length of time spent introducing and discussing the definition of the "minimally competent" candidate (approximately 60 minutes)?

Response	Percentage	Count
About right	77.3%	17
Not enough time	4.6%	1
Too much time	18.2%	4
Total responses	**100.0%**	**22**

6. How clear were you about the definition of the "minimally competent" candidate for the MCCQE Part I as you began the task of setting a pass score on Day 2?

Response	Percentage	Count
Very clear	72.7%	16
Clear	27.3%	6
Somewhat clear	0%	0
Not clear	0%	0
Total responses	**100.0%**	**22**

7. How clear did you find the information regarding the overview of standard setting that was provided the morning of Day 1?

Response	Percentage	Count
Very clear	50.0%	11
Clear	36.4%	8
Somewhat clear	13.6%	3
Not clear	0%	0
Total responses	**100.0%**	**22**

8. What is your impression of the length of training time you received for setting a passing score on the MCCQE Part I?

Response	Percentage	Count
About right	86.4%	19
Not enough time	9.1%	2
Too much time	4.6%	1
Total responses	**100.0%**	**22**

9. What is your impression of the amount of training you received for using the Bookmark method?

Response	Percentage	Count
Very clear	68.2%	15
Clear	27.3%	6
Somewhat clear	0%	0
Not clear	4.6%	1
Total responses	**100.0%**	**22**

10. How did you find the practice session for applying the Bookmark method on the afternoon of Day 1?

Response	Percentage	Count
Very clear	77.3%	17
Clear	18.2%	4
Somewhat clear	0%	0
Not clear	4.5%	1
Total responses	**100.0%**	**22**

11. How would you rate your understanding of how to apply the Bookmark method during Round 1 of the exercise?

Response	Percentage	Count
Very clear	63.6%	14
Clear	31.8%	7
Somewhat clear	0%	0
Not clear	4.6%	1
Total responses	**100.0%**	**22**

12. How would you rate your understanding of how to apply the Bookmark method during Round 2 of the exercise?

Response	Percentage	Count
Very clear	81.8%	18
Clear	13.6%	3
Somewhat clear	4.6%	1
Not clear	0%	0
Total responses	**100.0%**	**22**

13. What is your overall evaluation of the training that was provided for setting a pass score on the MCCQE Part I?

Response	Percentage	Count
Excellent	77.3%	17
Very good	13.6%	3
Good	4.6%	1
Fair	4.6%	1
Poor	0%	0
Total responses	**100.0%**	**22**

14. How would the training for setting a passing score on the MCCQE Part I have been improved?

Response	Percentage	Count
No changes required	22.7%	5
Different articles. The ones provided were not helpful / more reference material about the Rasch method	9.1%	2
More panelists	4.6%	1
Not sure or no response	63.6%	14
Total responses	**100.0%**	**22**

15. What factors influenced your placement of the Bookmark on Day 2? Please select all that apply.

Response	Percentage	Count
Description of the minimally competent candidate	86.4%	19
My perception of the difficulty of the test items	63.6%	14
The test item statistics (i.e., PR67)	59.1%	13
My experience with candidates in the field	50.0%	11
Knowledge and skills measured by the test items	31.8%	7
The impact data presented	59.1%	13
Panelist discussion	54.6%	12
Bookmark placement of other panelists	27.3%	6
Other: politics and expectations	4.6%	1
Total number of respondents	**100.0%**	**22**

16. How would you judge the length of time provided?

Response		Percentage	Count
About right		77.3%	17
Not enough time		0%	0
Too much time		22.7%	5
	Total responses	**100.0%**	**22**

17. Overall, how did you feel about participating in group discussions?

Response		Percentage	Count
Very comfortable		86.4%	19
Somewhat comfortable		9.1%	2
Unsure		0%	0
Somewhat uncomfortable		4.6%	1
Very uncomfortable		0%	0
	Total responses	**100.0%**	**22**

18. How did you find the impact data and discussions facilitate the panel to arrive at a passing score?

Response		Percentage	Count
Very helpful		77.3%	17
Helpful		18.2%	4
Somewhat helpful		0%	0
Not at all helpful		4.6%	1
	Total responses	**100.0%**	**22**

19. What level of confidence do you have in the final recommended passing score?

Response		Percentage	Count
Very confident		68.2%	15
Confident		27.3%	6
Somewhat confident		4.6%	1
Not at all confident		0%	0
	Total responses	**100.0%**	**22**

20. Please provide any additional comments or suggestions about the setting of a passing score on the MCCQE Part I.

	Response
1.	The exercise to set a pass was interesting and mind blowing. The training phase was educational and is crucial for all judges from all levels of medicine and surgery to understand the Bookmark method. In my case, I understood the principle only by Round 2.
2.	Apply other methods to compare? (mean & SD?) Also remind us that we can apply group 2 MOC credits for curriculum - exam developments. Respondent wrote on Q7: new to me - but articles distributed beforehand were helpful.
3.	Email support friendly & helpful. Availability of speakers between sessions also helpful.
4.	Too much time waiting around.
5.	Although at first I did not understand how the calculation was made, Fang explain well how the calculation is done and able to understand to arrive the right cut score. Excellent coordinators, facilitators and the MCC gang.
6.	Interesting process. Well done. Enjoyable process.
7.	Thank you! A very interesting & informative experience!
8.	Well organized, enthusiastic & friendly staff.
9.	I very appreciated learning how a cut score can be determined. I understand that all organizations cannot do it this way but it is a very robust example. Thanks!
10.	Very good meeting. Thank you for everything.
11.	Thanks a lot. See you again in the future.
12.	Some of the content needs to be reviewed. Thank you.
13.	Thank you for making this such an enjoyable and educational experience!

Note: Percentages may not total 100 per cent due to rounding.