MEDICAL COUNCIL OF CANADA | LE CONSEIL MÉDICAL DU CANADA

Technical Report on the Standard Setting Exercise for the Medical Council of Canada Evaluating Examination

Psychometrics and Assessment Services

May 2017

Table of Contents

1. BACKGROUND AND PURPOSE 4

2. PROCEDURES 4

2.1 Selecting a standard setting method 4

2.1.1 Bookmark method 5

2.1.2 Hofstee method 6

2.2 Selecting and assigning standard setting judges into two subpanels 6

2.3 Preparing materials for the standard-setting exercise 8

2.3.1 Test form 8

2.3.2 RP67 8

2.3.3 Ordered item booklet (OIB) 8

2.3.4 Item map 8

2.3.5 Practice Booklet and Practice OIB 9

2.4 Advance mailing 9

2.5 Activities during the two-day session 9

2.5.1 Day 1 – Training and practice 9

2.5.1.1 Familiarizing judges with the MCCEE 9

2.5.1.2 Defining the minimally competent candidate 9

2.5.1.3 Practice using the Bookmark method 10

2.5.2 Day 2 – Standard setting exercise 11

2.5.2.1 Round 1 (preliminary round) 11

2.5.2.2 Round 2 (final round) 12

2.5.2.3 Calculation of the Bookmark cut score 12

3. RESULTS 12

3.1 Bookmark results 12

3.2 Impact data 13

3.3 Hofstee results 14

3.4 Final recommended pass score 15

3.5 Post-session survey 15

4. CONCLUSIONS 16

5. REFERENCES 18

APPENDIX A: INVITATION LETTER AND DEMOGRAHIC SHEET 19

APPENDIX B: STANDARD SETTING MEETING AGENDA 23

APPENDIX C: DEFINITION OF THE MINIMALLY COMPETENT CANDIDATE 25

APPENDIX D: BOOKMARK FORM 26

APPENDIX E: HOFSTEE METHOD 27

APPENDIX F: SUMMARY OF RESPONSES TO POST-MEETING SURVEY 28

LIST OF TABLES

Table 1: Demographic Information by Standard Setting Subpanel 7
Table 2: Bookmark Cut Scores 13
Table 3: Failure Rates by Round and Candidate Cohort 13
Table 4: Summary of Hofstee Results by Round and Subpanel 14

LIST OF FIGURES

Figure 1: Hofstee Results and Impact 15

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 that is used to make classification decisions (e.g., pass/fail, grant/withhold a credential, award/deny a licence).

The Medical Council of Canada (MCC) Evaluating Examination (MCCEE) is a general assessment of the basic medical knowledge in the principal disciplines of medicine at the level of new medical graduates who are about to enter the first year of postgraduate training in Canada. It includes 180 multiple-choice questions (MCQ), of which 150 are scored and 30 are non-scored pretest or "pilot" questions. It is a four-hour, computer-based examination offered in both English and French in over 80 countries worldwide. International medical graduates (IMG), including American osteopathic graduates, must take the MCCEE as an eligibility prerequisite 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), that is designed to assess the readiness of IMGs for entry into postgraduate residency training programs in Canada.

The current pass score for the MCCEE was established in 2011. It is best practice to review the standard and the pass score regularly (e.g., every three to five years) to ensure that they remain appropriate and reflect the current standard to practise competently in the profession to protect public interest given the evolvement of the exam and test-taker population, and to reflect advancements in medicine and medical education.

On November 17-18, 2016, a panel of 21 physicians from across Canada met at the MCC's office in Ottawa to participate in a standard-setting exercise for the MCCEE. Staff from the Psychometrics and Assessment Services (PAS) directorate, with support from staff in Evaluation Bureau (EB), facilitated the meeting. The purpose of the meeting was to arrive at a recommended pass score for subsequent consideration and approval by MCC's Evaluating Examination Composite Committee (EECC), a body that is responsible for overseeing the MCCEE including the development and maintenance of the 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 of a new pass score for the MCCEE.

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 a number of considerations.

- First, the MCCEE 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 is determined by comparing an individual candidate's performance to a performance standard regardless of the performance of other candidates. Therefore, a criterion-refereced method of standard setting such as the Bookmark method is most appropriate for the MCCEE.
- Secondly, the MCCEE uses an MCQ format where candidates are asked to select one correct answer from five possible answers for each question (also referred to as item). We typically use a test-centered standard-setting method for MCQ exams (e.g., Bookmark or Angoff method) where expert judges review test items and provide judgments as to the adequate level of performance on those items. Conversely, we typically use an examinee-centered approach for performance exams (e.g., Borderline Group or Contrasting Group method) where judges review the performance of an actual group of examinees and provide judgments as to the adequate level of performance (Kane, 1998). The Bookmark method is a test-centered, criterion-referenced method that has been widely used for setting standards on licensure and certification examinations.
- Thirdly, using the Bookmark method, it is convenient to connect the judgment task of setting a cut score to the one-parameter (1PL) item response theory (IRT) model used to calibrate items and assemble test forms for the MCCEE. The 1PL IRT 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 IRT ability scale. In this way, candidates with scale 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, the MCQ component of which is similar to the MCCEE.

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 presented, one item 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 2.5.1.2). 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 a 2/3 odds. 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% chance of answering

items correctly beyond that point. 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% chance of answering an item correctly. However, their chance would likely fall below 0.67 probability or 2/3 odds.

Each individual panelist's cut score corresponds to the 1PL 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. This process can be repeated in two or three rounds. Impact data (i.e., pass/fail rate as a result 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 a holistic judgment on an acceptable cut score (criterion-referenced) and 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 lowest cut score that would be acceptable, even if *no* candidate attained that score?
- What is the highest cut score that would be acceptable, even if *every* candidate attained that score?
- What is the maximum tolerable failure rate?
- What is the minimum tolerable failure rate?

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 along with the cumulative percentage of candidates who would fail at each point along the score scale in an effort to define a cut score (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 and assigning standard setting panelists into two subpanels

Selecting a panel of well-qualified panelists is an important step to promote the validity of a standard-setting process and the resulting cut score. In view of 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 January 2016, the MCC, through various MCCEE test committees and EECC members, 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 43 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 21 participants (we originally selected 22 but one withdrew) and assigned them to two subpanels that were matched as closely as possible on key demographic variables, including: (1) gender, (2) geographic region, (3) ethnic background, (4) medical specialty and, (5) number of years in practice. The main purpose of using two subpanels was to assess the replicability of the cut score across two parallel but independent groups of physicians; a critical source of validity evidence in support 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 1 summarizes the demographic composition of the two subpanels.

Table 1: Demographic Information by Standard Setting Subpanel

Variable of Interest	Group	Subpanel 1	Subpanel 2	Total
Gender	Male	4	4	8
	Female	7	6	13
Geographic Region	West	1	1	2
	Prairies	2	2	4
	Ontario	5	5	10
	Quebec	1	1	2
	Maritimes	2	1	3
Ethnic Background	Caucasian	6	6	12
	Asian	2	2	4
	Other	3	2	5
Specialty	Primary Care	5	6	11
	Other Care	5	5	10
Number of Years in Practice Post-Residency	1-10	6	5	11
	11-20	3	2	5
	21-30	2	3	5

2.3 Preparing materials for the standard-setting exercise

2.3.1 Test form

The MCCEE is administered using a computer-based, linear-on-the-fly-test (LOFT) model. With the LOFT design, a unique test form is assembled in real-time by selecting items from a large pool of operational items each time a candidate takes the exam. Consequently, there are as many test forms as the number of candidates. Although each candidate receives a unique set of items, scores from all test forms are comparable as all forms are assembled to meet the same psychometric and content specifications and all items in the pool are pre-calibrated using the 1PL IRT model and linked to a common scale established for the item bank. Essentially, any test form can be used for the purpose of standard setting. From hundreds of test forms delivered in the most recent administration (i.e., September 2016), we selected a form that is typical and optimal in terms of meeting content and psychometric specifications. As any other form, this form consisted of 150 scored MCQs and 30 non-scored pilot items. Only the 150 scored items were used for standard setting.

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 0.67 response probability of answering an item correctly was calculated using the formula (Cizek & Bunch, 2007):

$$\theta_i = \beta_j + .693$$

where θ_i and β_j represent examinee ability and item difficulty, respectively. The RP67 values were calculated for each of the 150 items selected for the standard-setting exercise.

2.3.3 Ordered item booklet (OIB)

The 150 items were assembled into a 150-page OIB, one item per page, ordered from the easiest to the most difficult based on item difficulty parameters. For each item on each page, we also provided its item ID, answer key and ability required for a 2/3 chance to answer correctly (i.e., RP67).

2.3.4 Item map

An item map was prepared that included information about the item order (i.e., page number of each item) in the Standard Setting OIB, item ID, answer key, RP67 value, *p*-value (proportion of candidates who answered the item correctly) and content classification for each item.

2.3.5 Practice Booklet and Practice OIB

We also selected 50 items (approximately 1/3 of items required for each content area in a test form) and assembled a mini test form to be used for training panelists. We deliberately chose the items to represent a range of difficulty levels. A Practice Booklet was prepared by ordering the 50 items randomly. A Practice OIB was also prepared by ordering the same items 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.

2.4 Advance mailing

To assist panelists in preparing for the standard setting exercise prior to the meeting, we emailed in advance the meeting agenda and two research papers (De Champlain, 2013; Karantonis & Sireci, 2006) that provided overviews of standard setting and the Bookmark method. Panelists were asked to read the papers to gain a preliminary understanding of standard setting and the Bookmark method.

2.5 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.

2.5.1 Day 1 – Training and practice

The success of any standard setting exercise relies heavily on extensive training of standard setting panelists. To this end, we devoted Day 1 exclusively to training 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 EECC for consideration and approval. We then provided an overview of the MCCEE 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.

2.5.1.1 Familiarizing judges with the MCCEE

To familiarize the judges with the type of questions and difficulty level of the MCCEE, we gave them an hour to review the Practice Booklet (see section 2.3.5) and answer the 50 sample questions that were presented in random order. We then provided the panelists with an answer key to self-score their answers without sharing their resulting score with other judges. We followed this by a discussion of their perceived difficulty level of the questions and the range of content coverage, given the purpose of the MCCEE and its target test-taker population.

2.5.1.2 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 MCCEE, the target is a minimally-competent candidate for entry into residency training in Canada. On the afternoon of Day 1, we devoted an hour for panelists to discuss the definition of the minimally-competent candidate that

was used for standard setting for the MCCQE Part I. Both the MCCEE and MCCQE Part I are designed to assess the knowledge and skills required at the level of a new medical graduate who is about the enter the first year of supervised postgraduate training, even though the intended uses and test-taker populations are different. The definition was reviewed by the MCC's Chief Medical Education Advisor prior to the meeting and was deemed appropriate for the MCCEE.

We used the draft definition as a starting point for discussion. We asked panelists to share their thoughts, envision some minimally-competent candidates, discuss their characteristics, what they know or are capable of doing, things they may have difficulty in doing, what distinguishes "minimally-competent" from "non-competent, etc. The intention was to help the panelists converge on a more unified conceptualization of the minimally-competent candidate given the purpose of the MCCEE as well as get closer to defining a meaningful cut score. In the end, panelists only made a very slight change to the draft definition. The final definition is presented in Appendix C. We asked panelists to keep in mind the definition and the image of the minimally-competent candidate consistently throughout the two-day exercise.

2.5.1.3 Practice using the Bookmark method

After panelists familiarized themselves with MCCEE 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 in the afternoon of Day 1 prior to engaging in the actual full-scale standard-setting exercise on Day 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 50 items in the Practice OIB. The 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 the 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 at the page at which they felt a minimally-competent candidate would have at least a 0.67 probability or a 2/3 chance of correctly answering all items up to that point, with their chances decreasing beyond that page. We asked them to record their bookmark page on a Bookmark Form (see Appendix D). 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, *p*-value and item content classification.

To illustrate how a cut score is derived, we calculated the practice cut scores based on their bookmarked pages by individual judges, subpanel and full panel (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 how we eventually arrive at a final cut score.

Through training, hands-on practice and thorough discussions, panelists developed a very good understanding of the purpose, content and difficulty level of the MCCEE, the definition of the minimally-competent candidate, the standard setting process and the Bookmark method by the end of Day 1.

At the end of the day, we briefly introduced the Hofstee method and how it can be used to check the practicality of a standard set by the Bookmark method.

2.5.2 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.

2.5.2.1 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 an OIB containing 150 items for standard setting, ordered by difficulty from the easiest to the most difficult (see section 2.3.3). We also provided an item map to panelists that included information about the item order, item ID, answer key, RP67, *p*-value and item content classification. We instructed panelists to review the items in the OIB in the order presented, starting from page 1 and place a bookmark at the point at which they felt a minimally-competent candidate would have a 0.67 probability or 2/3 chances to correctly answer all items up to that page, with their chances decreasing beyond that point. 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, we projected the definition of the minimally-competent candidate on a screen in each room to allow panelists to refer to the definition at all times while making judgement on each item.

After panelists completed their Bookmark judgments, we asked them to provide and record answers to the four Hofstee questions as described in section 2.1.2 using a form (see Appendix E). Specifically, we asked panelists to specify the highest and lowest cut scores as well as the highest and lowest failure rates they believed would be reasonable for the MCCEE based on their holistic judgment.

We collected the completed Bookmark and Hofstee forms and during the panelists' lunch break, PAS staff tallied the bookmarks and calculated cut scores by individual panelist, subpanel and full panel. We also calculated the impact of the full panel's cut score on failure rate using candidate performance data on the MCCEE from 2015, as well as Hofstee results for each subpanel and the full panel.

We then presented the results and impact data from Round 1 to all panelists before splitting them once again into two groups in differing 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 2014 and 2015, based on the current cut score that was established in 2011.

The Round 1 exercise provided judges with an opportunity for a realistic practice and 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, panellists were better prepared for Round 2, that is the final round.

2.5.2.2 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., item 1); 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). We instructed the panelists to either place a second bookmark if they changed their mind after Round 1 discussions, or keep their first bookmark if so desired. 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. During a break, we gathered the forms, 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 first-time test-takers from 2015 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 2014 and 2015. 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 MCCEE candidates.

2.5.2.3 Calculation of the 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 0.67 probability of a correct response as expressed on an IRT θ scale). We used the median across panelists in each subpanel to obtain a subpanel's cut score. 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

In Table 2 we present a summary of the Bookmark cut scores for each subpanel and the full panel. Subpanel 1 had a slightly wider range and more variability across panelists in Round 1. As anticipated, the two subpanels converged in Round 2. Subpanel 2 converged slightly closer in Round 2, but overall, the range of cut scores was similar between the two subpanels. 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% confidence intervals around their cut scores using SEJ for each subpanel. The 95% intervals for Subpanel 1 (-0.73, -0.25) and Subpanel 2 (-0.67, -0.31) indicated very similar ranges and significant overlap between the two subpanels.

The θ cut score of -0.490 derived from Round 2 became the standard-setting panel's final recommended cut score. A θ score of -0.490 translates to a scale score of 261 on the MCCEE reporting scale of 50-500. It is slightly higher than the current θ cut score of -0.704 (or 250 on the reporting scale).

Table 2: Bookmark Cut Scores

		Cut Score (θ)	Min	Max	SD	SEJ
Round 1	Subpanel 1	-0.490	-0.958	0.174	0.61	0.18
	Subpanel 2	-0.490	-0.671	0.232	0.51	0.16
	Full panel	-0.490				
Round 2	Subpanel 1	-0.490	-1.495	0.435	0.39	0.12
	Subpanel 2	-0.490	-1.324	0.439	0.28	0.09
	Full panel	-0.490				
Final Cut Score		**-0.490**				

3.2 Impact data

As indicated earlier, we computed the impact of cut scores on failure rates using performance data from 3147 first-time test takers in 2015. These results are presented in Table 3. As the same cut score was arrived at in the two rounds of the standard-setting exercise, the resulting failure rates were the same. For comparison, historical failure rates are also shown for 2014-2015 based on the current cut score. As shown, the recommended θ pass score of -0.490 (261 on reporting scale) would result in 7% higher failure rate based on 2015 candidate performance data.

Table 3: Failure Rates by Round and Candidate Cohort

	Recommended Cut Score		First Time Test-Takers	All Test-Takers
	θ	Reported		
Round 1	-0.490	261	30%	38%
Round 2 (Final)	-0.490	261	30%	38%
Current Cut Score				
2015	-0.704	250	23%	30%
2014			24%	30%

3.3 Hofstee results

Table 4 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 4: Summary of Hofstee Results by Round and Subpanel

	Statistic	Subpanel 1	Subpanel 2	Full Panel
Round 1	Min Acceptable Percentage Cut Score	43.73	41.30	42.00
	Max Acceptable Percentage Cut Score	73.45	69.30	71.48
	Min Acceptable Failure Rate	11.64	18.50	15.00
	Max Acceptable Failure Rate	35.00	41.50	38.00
Round 2	Min Acceptable Percentage Cut Score	43.64	39.50	**42.00**
	Max Acceptable Percentage Cut Score	71.18	69.00	**71.21**
	Min Acceptable Failure Rate	13.64	19.00	**17.00**
	Max Acceptable Failure Rate	40.45	41.00	**41.00**

In Figure 1, the average answers from the full panel in Round 2 (as reported in Table 4) are plotted against a cumulative percentage of candidates who would fail at each point along the θ ability scale using the performance data of first-time test-takers from 2015 MCCEE administrations. Panelists felt that the cut score should be no lower than 42% and no higher than 71%. Similarly, they indicated that the failure rate should be at least 17% but no higher than 41%. The coordinates (max. cut score and min. failure rate) and (min. cut score and max. failure rate) are linked by a red, dotted line. The point of intersection between this line and the cumulative frequency distribution corresponds to a cut score of close to -0.49 if we drew a vertical line down from this point to the horizontal axis. This happens to be the cut score set by the Bookmark method which would result in approximately a 30% failure rate for 2015 first-time test-takers. As indicated earlier, the Hofstee method was not our primary method for setting the standard for the MCCEE; 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 what the cut score and failure rate should be from a political perspective.

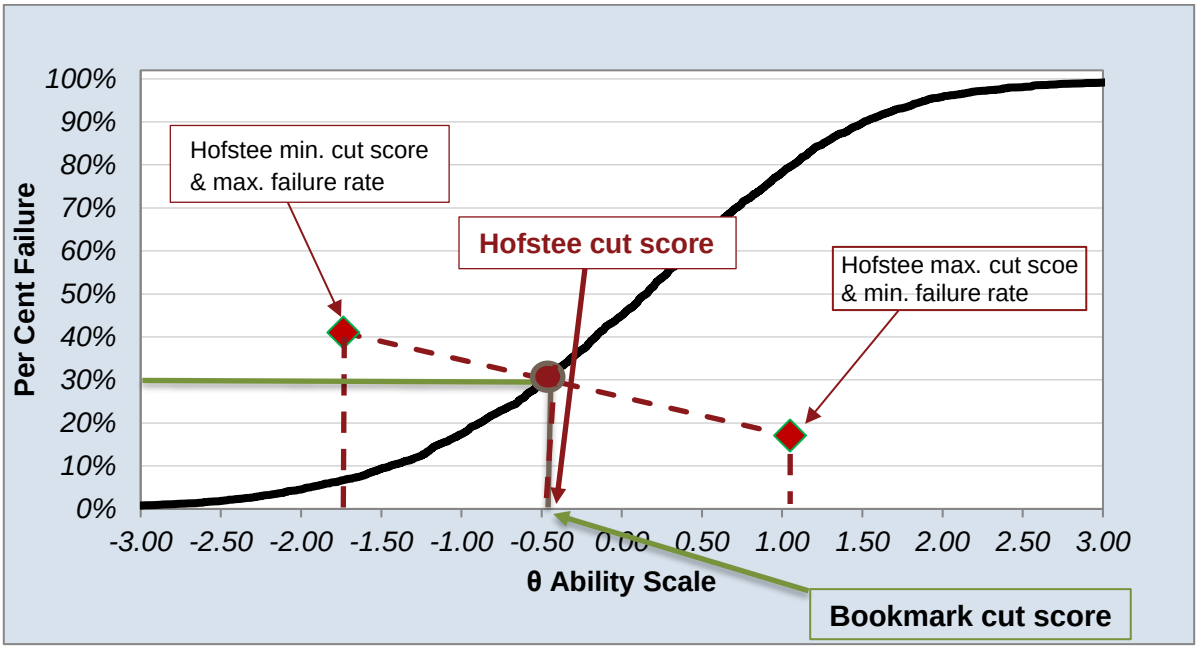

Figure 1: Hofstee Results and Impact

3.4 Final recommended pass score

After a rigorous, two-day standard-setting exercise, the panel of 21 physicians recommended a new pass score of -0.490 on the θ scale (261 on reporting scale) that was subsequently brought forward to the EECC for consideration and approval.

3.5 Post-session survey

At the conclusion of the meeting, we provided panellists an opportunity to provide feedback on the standard-setting exercise by answering a survey anonymously. Full results of the survey are presented in Appendix F. Eighteen 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 MCCEE and standard setting. We surveyed panelists about the clarity of the information provided. All respondents found the information regarding the MCCEE to be very clear (66.7%) or clear (33.3%). Respondents also found the information regarding standard setting to be very clear (66.7%), clear (28.6%) or somewhat clear (4.8%).
- Central to the standard setting exercise is the notion of the minimally-competent candidate. Most panelists felt they benefited from the discussion of the "minimally-competent" candidate and they found the discussion very helpful (61.9%), helpful (23.8%) or somewhat helpful (14.3%). About 95% of respondents considered the time spent on the discussion adequate. One person felt that too much time was spent on this issue. Almost all respondents felt they were very clear (52.4%) or clear (42.9%) about the definition of the "minimally-competent" candidate as they began the task of setting a cut score.
- 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. Overall, respondents felt that the training was excellent

(70%), very good (25%) or good (5%). All respondents indicated the amount of training they received for using the Bookmark method was very adequate (66.7%) or adequate (33.3%) and that the practice session for applying the Bookmark method was very helpful (95.2%) or helpful (4.8%).

- We asked participants about their level of understanding of how to apply the Bookmark method. Respondents felt they had very good (76.2%) or good (23.8%) understanding during Round 1 of the exercise. Their understanding improved after Round 1 as 85.7% indicated very good and 14.3% indicated good understanding of the method during Round 2 of the exercise.
- We solicited panellists' opinions on factors that influenced their placement of their bookmark. Multiple factors were considered from the most used to the least used: definition of the minimally-competent candidate, their perception of the difficulty of the test items, their experience with the candidates in the field, knowledge and skills measured by the test items, judge discussions, bookmark placement of other judges, the impact data presented and item statistics. It is worth noting that 100% of respondents considered the definition of the minimally-competent candidate when making judgment on items, as they were trained and instructed to do.
- 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 (55%), helpful (35%) or somewhat helpful (10%) in facilitating the panel to arrive at a defensible pass score.
- Finally, and most importantly, panelists indicated they were very confident (80%) or confident (20%) 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 two subpanels independently arrived at the same cut score in Round 1; there was absolutely no influence from each other. This occurred again in Round 2 though it is possible that by this time, they might have been influenced by Round 1 results, impact data and discussions with other panelists. This provides evidence to support the balanced assignment of the two subpanels as well as successful training to calibrate judges to a common understanding of the process.
2. Both rounds resulted in the same cut score at the subpanel and full panel level with some individual panelists having adjusted their cut score after Round 1. This indicates that the training provided to panelists in Round 1 and Round 2 further reinforced their understanding of the standard setting process.
3. The 95% intervals around the cut score constructed using the standard error of judgment for Subpanel 1 (-0.73, -0.25) and Sub-panel 2 (-0.67, -0.31) indicate very similar ranges and significant overlapping between the two subpanels. This provides evidence to support the generalizability of the cut score if different panels were used to follow the same process.

4. The recommended Bookmark cut score was within the acceptable range defined by the Hofstee method based on judges' holistic judgment. As a matter of fact, the cut score defined by the Hofstee method overlaps perfectly with the Bookmark cut score. This indicates that the criterion-referenced cut score derived using the Bookmark method is realistic and consistent with political considerations.
5. The results of the post-session survey indicate a very positive experience from the panelists' point of view and confidence in the training provided.
6. Panelists expressed high confidence in the standard-setting process and the final recommended pass score as indicated by the post-session survey results.

All of 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 political perspectives.

The recommended pass score was presented to the EECC on December 8, 2016, along with an overview of the standard setting process, followed by the impact data. The EECC unanimously approved the recommended θ pass score of -0.490 (261 on reporting scale) for the MCCEE. The new pass score will be implemented as of May 2017.

5. References

Cizek, G. J. (2012). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, Methods, and Innovations* (pp. 3-14). New York, NY: Routledge.

Cizek, G. J. and Bunch, M. B. (2007). Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests (pp.155-189). 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.

De Champlain, A. F. (2004). Ensuring that the competent are truly competent: An overview of common methods and procedures used to set standards on high-stakes examinations. *Journal of Veterinary Medical Education*, 31, 61-5.

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.

Kane, M. (1994). Validating the Performance Standards Associated With Passing Scores. In *Review of Educational Research*. Fall 1994 64 (3), 425-461.

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

Appendix A: Invitation Letter and Demographic Sheet

Invitation Letter

Dear Doctor,

The purpose of this letter is to invite you to express interest in serving as a potential panelist in a standard setting exercise for the Medical Council of Canada (MCC) Evaluating Examination (MCCEE). The results of this exercise will provide the MCC with valuable information in determining the passing standard that will be used for the MCCEE in the years to come. We are hoping to secure your participation for this exercise, which is scheduled to take place on **November 17-18, 2016**, at the MCC's head office in Ottawa, ON. Staff from the MCC's Psychometrics and Assessment Services department will be moderating this standard setting exercise with staff support from the Evaluation Bureau. We are sending out this notice to solicit volunteers from which we will assemble panels, based on demographic details and other considerations that we are asking you to kindly provide in the attached questionnaire. Participants will be selected to reflect diverse medical specialties and practice contexts across Canada.

Panelists will be trained to evaluate examination materials and will be guided through a set of procedures to set the passing score in small working groups. Invited panelists will be provided an honorarium of $600 per day (2-day meeting) plus reasonable travel expenses to Ottawa, ON. Panelists will also be issued a certificate of participation which they can use to apply for CME credits.

If you are interested in participating, we ask that you fill out the enclosed questionnaire and return it to the MCC. We also ask that you block the above-mentioned November 17 and 18, 2016 dates in your calendar. **Finally, we kindly request that you respond by April 28, 2016** with the completed demographic sheet and we will confirm your participation by April 27, 2016.

Thank you very much for your interest and support.

Sincerely,

Demographic Information Sheet

The information requested below is being collected to help the Medical Council of Canada (MCC) obtain a pan-Canadian representative panel to recommend a passing score on the MCC Evaluating Examination (MCCEE) (*http://mcc.ca/examinations/mccee/*). 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 in Ottawa on **November 17 and 18, 2016**, therefore we are asking panelists to be available on both days until **5:00 p.m.**

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

Medical Council of Canada
100-2283 St-Laurent Blvd.
Ottawa, ON K1G 5A2

Name (please print): _____

Preferred contact information (mailing address, email address & phone number):

_____

---

1. Number of years in practice post-residency:

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

2. Number of years' experience supervising residents:

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

3. Do you have experience supervising International Medical Graduates in Canada?

No	☐
Yes	☐

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

No	☐
Yes	☐

4b. If yes, which test committee?

MCC Evaluating Examination ☐
MCC Qualifying Examination Part I ☐
MCC Qualifying Examination Part II ☐
NAC Examination ☐

Note: Being a test committee member is not a requirement to participate in this exercise.

5. Country of postgraduate medical training:

Canada ☐
Other _____ ☐

6. Region of the country in which you live:

Alberta ☐
British Columbia ☐
Manitoba ☐
Maritimes ☐
Ontario ☐
Quebec ☐
Saskatchewan ☐
Territories ☐

7. First Language:

English ☐
French ☐
Other (_____) ☐

8. Gender:

Female ☐
Male ☐

9. Ethnicity:

Asian ☐
Black ☐
Caucasian ☐
First Nations ☐
Hispanic ☐
Other ☐

10. Medical Specialty:

Family Medicine ☐
Internal Medicine ☐
Obstetrics and Gynecology ☐
Pediatrics ☐
PHELO ☐
Psychiatry ☐
Surgery ☐
Other _____ ☐

11. Type of community in which you work:

Rural ☐
Urban ☐

12. Type of care setting:

Community-based ☐
Hospital-based ☐

Appendix B: Standard Setting Meeting Agenda

MEDICAL COUNCIL OF CANADA
Evaluation Examination (EE) Standard Setting Exercise
November 17-18, 2016
Medical Council of Canada Office, University Boardroom
2283 St. Laurent Blvd., Suite 100, Ottawa, ON

DAY 1 – November 17, 2016

TIME	ACTIVITIES
08:00	CONTINENTAL BREAKFAST
08:30	Welcome and introductions
08:45	Review agenda and objectives
09:00	Overview of MCCEE
09:15	Overview of standard setting
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
14:30	BREAK
14:45	Panelists practise Bookmark method using the Ordered Item Booklet (OIB) for Practice Test (subpanels in two rooms)
15:45	Post-Bookmark training discussion and clarification
16:15	Wrap-up/Overview of Day 2
16:30	End of Day 1

DAY 2 – November 18, 2016

TIME	ACTIVITIES
08:00	CONTINENTAL BREAKFAST
08:30	Round 1: Split into two subpanels and two rooms • Panelists independently provide Bookmark and Hofstee judgements using the OIB for standard setting
11:30	LUNCH
	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 full panel discussion
13:30	Round 2: Subpanels in two rooms • Panelists independently provide Bookmark and Hofstee judgements using the OIB for standard setting
15:00	BREAK
	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: Definition of the Minimally Competent Candidate

The “minimally competent” candidate entering supervised practice is a candidate who possesses the minimum level of knowledge and skills required to safely practice medicine under supervision. In contrast to a non-competent candidate, the “minimally competent” candidate’s performance is acceptable, despite gaps in their knowledge and clinical decision-making skills.

Appendix D: Bookmark Form

Standard Setting for the MCC Evaluating Examination
The Bookmark Method

Subpanel: _____ **Panelist Name:** _____

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: Hofstee Method

Standard Setting for the MCC Evaluating Examination
The Hofstee Method

Subpanel: _____ **Panelist Name:** _____

Please answer the following questions at the end of each Round.

1. Consider the content as a whole, what is the **highest** percent correct pass score that would be acceptable, even if every candidate attained that score?

   Round 1: _____ Round 2: _____

2. Consider the content as a whole, what is the **lowest** percent correct pass score that would be acceptable, even if no candidate attained that score?

   Round 1: _____ Round 2: _____

3. What is the **maximum** acceptable failure rate?

   Round 1: _____ Round 2: _____

4. What is the **minimum** acceptable failure rate?

   Round 1: _____ Round 2: _____

Appendix F: Summary of Responses to Post-Meeting Survey

1. Which panel did you participate in?

Response	Chart	Percentage	Count
Panel 1 (University room)		52.4%	11
Panel 2 (Barr/Berard room)		47.6%	10
		Total Responses	**21**

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

Response	Chart	Percentage	Count
Very clear		66.7%	14
Clear		33.3%	7
Somewhat clear		0.0%	0
Not clear		0.0%	0
		Total Responses	**21**

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

Response	Chart	Percentage	Count
Very clear		66.7%	14
Clear		28.6%	6
Somewhat clear		4.8%	1
Not clear		0.0%	0
		Total Responses	**21**

4. How did you find the discussion of the "minimally competent" candidate for the MCCEE during the training on the afternoon of Day 1?

Response	Chart	Percentage	Count
Very helpful		61.9%	13
Helpful		23.8%	5
Somewhat helpful		14.3%	3
Not helpful at all		0.0%	0
		Total Responses	**21**

5. How would you judge the length of time spent (approximately 1 hour) on the afternoon of Day 1 to discuss the definition of the "minimally competent" candidate?

Response	Chart	Percentage	Count
About right		95.0%	19
Too little time		0.0%	0
Too much time		5.0%	1
		Total Responses	**20**

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

Response	Chart	Percentage	Count
Very clear		52.4%	11
Clear		42.9%	9
Somewhat clear		4.8%	1
Not clear		0.0%	0
		Total Responses	**21**

7. What is your impression of the amount of training you received for using the Bookmark Method?

Response	Chart	Percentage	Count
Very adequate		66.7%	14
Adequate		33.3%	7
Somewhat adequate		0.0%	0
Not adequate		0.0%	0
		Total Responses	**21**

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

Response	Chart	Percentage	Count
Very helpful		95.2%	20
Helpful		4.8%	1
Somewhat helpful		0.0%	0
Not helpful at all		0.0%	0
		Total Responses	**21**

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

Response	Chart	Percentage	Count
Very good		76.2%	16
Good		23.8%	5
Fair		0.0%	0
Poor		0.0%	0
		Total Responses	**21**

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

Response	Chart	Percentage	Count
Very good		85.7%	18
Good		14.3%	3
Fair		0.0%	0
Poor		0.0%	0
		Total Responses	**21**

11. What is your overall evaluation of the training that was provided for setting a pass score on the MCCEE?

Response	Chart	Percentage	Count
Excellent		70.0%	14
Very good		25.0%	5
Good		5.0%	1
Fair		0.0%	0
Poor		0.0%	0
		Total Responses	**20**

12. What factors influenced your placement of Bookmark on Day 2? (Select all that apply.)

Response	Chart	Percentage	Count
Description of the minimally competent candidate		100.0%	20
My perception of the difficulty of the test items		95.0%	19
The test item statistics (i.e., RP67, p-value)		35.0%	7
My experience with candidates in the field		95.0%	19
Knowledge and skills measured by the test items		75.0%	15
The impact data presented		40.0%	8
Panelist discussions		70.0%	14
Bookmark placement of other panelists		25.0%	5
Other (please specify)		0.0%	0
		Total Responses	**20**

13. How did you find the impact data and discussions in facilitating the panel to arrive at a defensible pass score?

Response	Chart	Percentage	Count
Very helpful		55.0%	11
Helpful		35.0%	7
Somewhat helpful		10.0%	2
Not helpful at all		0.0%	0
		Total Responses	**20**

14. How confident do you feel in the final recommended pass score?

Response	Chart	Percentage	Count
Very confident		80.0%	16
Confident		20.0%	4
Somewhat confident		0.0%	0
Not at all confident		0.0%	0
		Total Responses	**20**