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 II

Psychometrics and Assessment Services

January 2019

Table of Contents

_Toc536190861
1. INTRODUCTION 2
2. PROCEDURES 3
2.1. Selecting a standard-setting method 3
2.1.1. Borderline group method 3
2.1.2. Hofstee method 4
2.2. Selecting and assigning panelists into two subpanels 4
2.3. Advanced mailing 5
2.4. Description of events during the three-day meeting 6
2.4.1. Training 6
2.4.1.1. Description of the just qualified candidate – Discussion 6
2.4.1.2. Standard setting and Borderline method – Training 7
2.4.2. Collection of ratings – Borderline group method 7
2.4.2.1. Data sources 7
2.4.2.2. Initial round 8
2.4.2.3. Final Round 8
2.4.2.4. Incorporating political and other considerations: The Hofstee method 9
2.4.3. Calculation of the pass score 9
2.4.4. Post-session survey 10
3. RESULTS 10
3.1. Borderline group results 10
3.2. Generalizability theory results 10
3.3. Impact data – Pass rates 12
3.4. Hofstee results 12
3.5. Summary of evaluation survey findings 13
4. CONCLUSIONS 14

REFERENCES 16

APPENDIX A: INVITATION LETTER AND DEMOGRAPHIC SHEET **17**
APPENDIX B: AGENDA **22**
APPENDIX C: PERFORMANCE LEVEL DESCRIPTORS MCCQE PART II **24**
APPENDIX D: HOFSTEE PAPER FORM **25**
APPENDIX E: SUMMARY OF RESPONSES TO POST-MEETING SURVEY **26**

1. Introduction

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 and operationalizing the resulting conceptual standard as a numerical pass score that is used to make classification decisions (e.g., pass/fail, grant/withhold a credential, award/deny a licence). A rigorous and valid process for standard setting should be adhered to for licensing examinations (Cizek, 2012). In this report, we outline the processes, procedures and results of a standard-setting exercise carried out for the Medical Council of Canada's Qualifying Examination (MCCQE) Part II.

The MCCQE Part II is a national, standardized examination that assesses the core abilities of candidates to apply medical knowledge, demonstrate clinical skills, develop investigational and therapeutic clinical plans, as well as demonstrate professional behaviours and attitudes at a level expected of a physician in independent practice in Canada. The MCCQE Part II is a performance assessment composed of a series of Objective Structured Clinical Examination (OSCE) stations that have two slightly different formats: (1) long stations (14 minutes) with checklist items, oral questions and rating scales and; (2) paired stations with checklist items, rating scales and extended match questions. Each encounter for the long station consists of encounters with Standardized Patients or Standardized Professionals (SPs). The paired stations consist of two six-minute components in any combination: an encounter component with an SP, a non-encounter component consisting of a reading task, or responding to one or more extended match questions. For example, there could be a station with two encounters or a station with a reading component then an encounter, or a station with an encounter and then extended match questions.

The pass score for the MCCQE Part II was last established in spring 2015. 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 examination (e.g., new blueprint, new format, etc.). This is to ensure that the standard and pass score remain appropriate and reflect the current standard to practise competently in the profession, to protect public interest and to reflect advancements in medicine and medical education.

From December 10 to 12, 2018, a panel of 20 physicians from across Canada met at the Medical Council of Canada (MCC) offices in Ottawa to participate in a standard-setting exercise for the MCCQE Part II. Staff from the Psychometrics and Assessment Services (PAS), 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 the Central Examination Committee (CEC), a body that is responsible for overseeing the MCCQE Part II, 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 three-day exercise that led to the recommendation of a new pass score for the MCCQE Part II.

2. Procedures

In this section, we present how the standard-setting method was selected, a description of how the panelists were selected, the information provided to the panelists prior to and during the three-day meeting, the method used to set the pass score, and a description of the events that took place during the three-day meeting.

2.1. SELECTING A STANDARD-SETTING METHOD

- First, the MCCQE Part II is a criterion-referenced exam for which a pass score should be defined as an acceptable amount of knowledge and skills 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-referenced method of standard setting such as the Borderline Group method is most appropriate for the MCCQE Part II.

- Secondly, the MCCQE Part II is a performance exam consisting of a series of OSCE stations. Examinee-centered standard-setting methods are most appropriate for performance assessments (e.g., Borderline Group or Contrasting Group method) where judges review the performance of a group of examinees and provide judgments as to the adequate level of performance (Cizek & Bunch, 2007). Examinee-centered methods are particularly well suited to the complex multidimensional nature of performance assessments. The Borderline Group method is an examinee-centered, criterion-referenced method that has been used for setting standards on licensure and certification examinations similar to the MCCQE Part II.

- We have used the Borderline Group method successfully for setting a standard on the MCCQE Part II in 2015 and the National Assessment Collaboration (NAC) Examination, which is similar to the MCCQE Part II, in 2013.

We also chose to complement the Borderline Group method with the Hofstee method. We describe the Borderline Group and Hofstee methods below.

2.1.1. Borderline group method

The Borderline Group method requires that panelists provide a holistic judgment of each candidate score sheet and assign each to one of the three levels (1 to 3), corresponding to *unacceptable/poor* (1), *just qualified/borderline pass* (2), or *acceptable/good* (3) performance on an OSCE station. A full description of how we implemented this method is outlined in the Standard Setting and Borderline Method – Training and Collection of Ratings – Borderline Group method sections below.

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 Borderline Group 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 along with the cumulative percentage of candidates who would fail at each point along the scale in an effort to define a pass score. The Hofstee method is usually not used as a standalone method. For our purpose, we used it to complement the Borderline Group method and provide a "reality check" on the pass score set using the Borderline Group method. A more detailed description of the Hofstee method is provided in Cizek & Bunch (2007) and Hofstee (1983).

2.2. SELECTING AND ASSIGNING PANELISTS INTO TWO SUBPANELS

Many features of a standard-setting exercise can influence the validity of the recommended pass score as well as its associated process. One of these features is the selection of well-qualified panelists. 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 July 2017, the MCC sent an email to physician Test Committee members and physician examiners soliciting participation in our standard-setting exercise. This solicitation resulted in more

2.3. ADVANCED MAILING

To assist panelists in preparing for the standard-setting exercise prior to the meeting, we emailed in advance the following documents: (1) an agenda for the meeting (see Appendix B); (2) a description of the *unacceptable/poor*, *just qualified/borderline pass*, and *acceptable/good* candidates generated by the OSCE Test Committee and reviewed by the CEC (see Appendix C) and; (3) three papers which provided an overview of standard setting (Boulet, De Champlain, McKinley, 2003; De Champlain, 2004; De Champlain, 2013).

2.4. DESCRIPTION OF EVENTS DURING THE THREE-DAY MEETING

The agenda for the three-day meeting is provided in Appendix B. The morning of the first day was devoted to training the panelists, followed by two rounds of collecting panelists' ratings over the remainder of the three-day meeting.

2.4.1. Training

The success of any standard-setting exercise relies heavily on extensive training of standard-setting panelists. To this end, we devoted the morning of Day 1 exclusively to training the panelists. The meeting began with an introduction of panelists and facilitators 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 II including its purpose, content, and station formats as described above, as well as scoring information.

We followed this with a thorough discussion of the *just qualified/borderline pass* candidate as described below. We wrapped up training with a training on the use of the Borderline Group method for the long and paired stations (the training of the paired station occurred prior to the first paired station on the second day of the exercise), which we also describe below.

2.4.1.1. Description of the just qualified candidate – Discussion

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 II, working with the OSCE test committee and the CEC, MCC staff members generated a description of *unacceptable/poor*, *just qualified/ borderline pass*, and *acceptable/good* candidates. These descriptors are presented in Appendix C. We reviewed the definitions and facilitated a group discussion to ensure that all panelists had a common understanding of candidate performance prior to the training sessions (approximately 45 minutes were devoted to this discussion). After discussion, we asked panelists to utilize these descriptions for the standard-setting exercise.

2.4.1.2. Standard setting and Borderline method – Training

After panelists reached a common understanding of the *just qualified* candidate, we provided a step-by-step training on how to use the Borderline Group method to recommend a pass score. Prior to commencing the collection of ratings for each station, we conducted a thorough training session, utilizing three videos for a long station (*good*, *borderline pass*, and *poor* performances) and two videos for a paired station (*good* and *borderline pass* performances). We selected different stations for training than the 10 stations used for the remainder of the exercise.

The purpose of the training sessions was to familiarize the panelists with the format of the stations as well as with *good* and *borderline pass* performances. For each training station (and subsequent 10 operational stations) we followed a four-step process:

1. A Test Development Officer (TDO) outlined the objective of the station.
2. A TDO reviewed the score sheet and score key.
3. The panelists reviewed two to three video performances.
4. The group discussed the video performances.

In addition, for the long station we had all the panelists practice entering ratings in our standard-setting tool. To ensure a common understanding of the categories of performance, and of the *just qualified/borderline pass* candidate in the context of the MCCQE Part II, we allowed ample time for discussion for each station type and each performance. Together, the training was approximately two and a half hours in length; one and a half hours for the long station and one hour for the paired station.

2.4.2. Collection of ratings – Borderline group method

Following training, we assigned the two panels to different rooms and a psychometrician facilitated each panel (Subpanel 1 and 2). For each station, a TDO for each subpanel followed the four-step process described above: (1) outlined the objective of the station, (2) reviewed the score sheet and score key, (3) had the panel review two to three video performances, and (4) facilitated a group discussion.

Subsequently, the panelists independently reviewed a set of 50 candidate score sheets for that station, ordered from the highest to the lowest station score, and assigned a rating from 1 to 3 (again, either *unacceptable/poor*, *just qualified/borderline pass*, or *acceptable/good*). There was no limit specified on the number of *borderline* candidates that they could identify. We then repeated the process for each station.

2.4.2.1. Data sources

We conducted the standard-setting exercise using the October 2018 test form of the MCCQE Part II and used stratified random sampling by total score to select 50

candidates whose performance represented a wide range of ability levels: (1) 34 per cent with a total score between 0 and 45, (2) 34 per cent with a total score between 45 and 65 (this range is the middle of the score distribution), and (3) 32 per cent with a total score between 65 and 100. Because watching candidate videos would be too time-consuming for 50 candidates per station for 10 OSCE stations, we used the actual candidate score sheets for each station as a proxy to candidate performance. The candidate score sheets were ordered from the highest to lowest station score for each station. For each candidate score sheet, each panelist provided a rating of 1 for *unacceptable/poor* performance; 2 for *just qualified/borderline pass* performance or; 3 for *acceptable/good* performance. In summary, for each of the two rounds, for each panelist, we collected 50 data points per station (one data point for each of the 50 candidate sheets) and 500 (50 x 10) ratings across the 10 stations.

2.4.2.2. Initial round

Initially, we gave panelists approximately 90 minutes to complete the rating task for the first few stations. Over the course of reviewing the 10 stations, we reduced these allotments to approximately 45 minutes, based on observed pacing. Panelists were always allowed more time if required, and each panelist provided ratings independently of other panelists. No discussion of ratings took place during this part of the exercise. Panelists entered their ratings electronically into an MCC-designed standard-setting electronic data capture tool. Panelists completed all ratings for the Initial Round by the end of the second day of the three-day exercise. We then asked the panelists to provide these judgments, as described in the Hofstee section below.

Before the beginning of the Final Round, we reconvened the two subpanels and presented the following information to both groups at the same time: (1) an explanation of how the pass score for each panelist was calculated (2) a description of the pass score by subpanel and combined across subpanels (3) the percentage of failures for Canadian Medical Graduates, Canadian Postgraduate first-time test takers (CMG-CPG 1st), by panelist, by subpanels and overall (4); the percentage of failures for first-time test takers, by panelist, by subpanel, and overall (5) Hofstee results, and (6) historical pass rates. We then separated the two subpanels to discuss the impact data for approximately 15 minutes. Each subpanel appointed a spokesperson to present a summary of their subpanel's discussion to the full group (approximately 10 minutes), which we then followed with a full panel group discussion (approximately 10 minutes).

2.4.2.3. Final Round

The meeting then proceeded with the collection of each panelist's independent judgments for each of the stations in the Final Round according to the following two-step process: (1) a brief summary of the content of each station and; (2) their second round of ratings of the 1-3 standard-setting judgments (*unacceptable/poor*, *just qualified/borderline*, or *acceptable/good*) for each of the 50 candidate score sheets. In

the MCC-designed standard-setting electronic data capture tool, ratings from the Initial Round were presented to panelists on the same screen for their reference. Following the Final Round, we collected the panelists' Hofstee data and a presentation of the Final Round MCCQE Part II pass score.

2.4.2.4. Incorporating political and other considerations: The Hofstee method

Prior to concluding each round, we asked panelists to answer four specific questions which define the Hofstee method as delineated above. The latter is generally viewed as a procedure which allows judges to gauge the appropriateness of standards considering a reality or reasonableness check that includes both criterion-referenced (acceptable pass score) and norm-referenced (acceptable failure rate) considerations. A description of the method was presented to the group followed by the entry of their judgments on paper (see Appendix D). Specifically, panelists were asked to specify the lowest and highest pass scores that they believed were reasonable for the MCCQE Part II. Additionally, panelists were asked to provide the lowest and highest failure rates that they felt were tolerable. Panelists provided acceptable low and high pass score values on the percent-correct scale (i.e., between 0 and 100).

Since Hofstee ratings permit the integration of both criterion- and norm- referenced considerations to gauge the appropriateness of pass scores derived using the Borderline Group method, our hope was that the pass scores would fall within the range of acceptable values as provided by the panelists (i.e. their "gut" estimates).

2.4.3. Calculation of the pass score

A panelist's pass score on an OSCE station corresponded to the median station score for those candidates they identified as *just qualified/borderline*. To illustrate, assume that Panelist A classified the following score sheets for Station 1 as j*ust qualified/borderline*: 66.5, 62.7, 65.8, 63.4, and 61.9. These values are the Station 1 scores associated with the five candidates that Panelist A judged as *just qualified/borderline*. Computing the median of these score sheets, 63.4, yields the estimate of the cut score for Panelist A for Station 1. We repeated this process for each station and for each panelist. Once we obtained the station cut scores for each panelist, we calculated the median of the panelist's 10 station cut scores as that panelist's overall MCCQE Part II cut score. Since panelists were organized into two subpanels, we calculated the median of the 10 panelists' pass scores and that was the pass score for that subpanel. Finally, we averaged the two estimates from the two subpanels to obtain an overall recommended MCCQE Part II pass score.

It is important to reiterate that throughout the three days, we routinely reminded panelists of the definition associated with the *just qualified/borderline* candidate and the purpose of the examination as they were carrying out the task of rating the candidate score sheets.

2.4.4. Post-session survey

The standard-setting exercise concluded by us asking all panelists to complete an evaluation survey which gauged their impressions of various aspects of the exercise as well as their confidence in the recommended pass score for the MCCQE Part II.

3. Results

3.1. BORDERLINE GROUP RESULTS

In Table 2 we present the computed pass scores for subpanel 1 and 2 as well as the mean of both panels for the Initial Round and the Final Round. As shown in Table 2, the Initial Round and the Final Round ratings were very similar across subpanels; however, the variability across raters decreased in the Final Round but the results were very similar across the initial and final rounds.

Table 2: Summary of pass scores by round and panel

	Initial round					
Statistic	N	Min.	Max.	Median	Mean	SD
Subpanel 1	10	48.4	58.9	**54.1**	54.3	3.6
Subpanel 2	10	49.5	59.1	**52.2**	53.6	3.3
Across panels	20			**53.2**		

	Final round					
Statistic	N	Min.	Max.	Median	Mean	SD
Subpanel 1	10	48.1	55.9	**53.2**	53.7	2.5
Subpanel 2	10	51.4	58.8	**53.8**	53.3	2.4
Across panels	20			**53.5**		

3.2. GENERALIZABILITY THEORY RESULTS

Generalizability (G) Theory is a statistical theory that provides a framework to estimate the dependability (i.e., reliability) of behavioural measurements (Shavelson & Webb, 1991). Dependability refers to the accuracy of generalizing from a person's observed score on a test or other measure to the average score that person would have received under all the possible conditions that the test user would be equally willing to accept (Shavelson & Webb, 1991). G-theory provides a summary coefficient reflecting the level of dependability (D-coefficient) and a generalizability coefficient (G-coefficient) that is analogous to classical test theory's reliability coefficient. Multiple sources (commonly called facets) of error in a measurement, can be estimated separately in a single analysis (e.g., persons or candidates, items, or in the case of OSCEs, stations,

raters and subpanel). The purpose of our analyses was to determine how much variance was attributable to sources that are undesirable, such as raters, subpanels, and stations and how much variance was due to actual differences in candidate abilities (true score variance, which is desirable in an effort to separate passing from failing candidates).

We conducted a G-study with three facets (*station*, *rater* and *subpanel*) in a *person* x *station* x (*rater: subpanel*) design. In other words, the same 50 candidates were rated on the same *stations* by *panelists* who were nested (assigned) to a specific *subpanel*. We used the ratings obtained from the Final Round for these analyses. In Table 3, we show the variance components for the candidates' ratings as well as each source of possible measurement error. The largest facet, not surprisingly, was the *person* x *station* interaction which accounted for 53.3 per cent of the total variance. This indicates that the performance of candidates (on the 1 to 3 scale) varied by station. This is commonly referred to as case specificity (Norman, Bordage, Page & Keane, 2006), which implies that success on any case or station is specific to that case and does not generalize very well to other stations. This is a common occurrence in OSCEs due to the smaller number of stations that can be realistically administered in an exam form (as compared to Multiple-Choice Questions, for example). The second largest effect was the *person* facet (22.6 per cent of total variance), which indicates that candidates did differ in their overall ability. This is akin to true score variance and suggests that the ratings for setting the pass score for the MCCQE II was able to separate out candidates, in terms of their ability level. The third largest effect was reported for the *station* facet which accounted for 5.0 per cent of the total score variance. This suggests that the ratings for setting the pass score on the stations differed, therefore the resulting pass score would change slightly if a different set of stations were used in subsequent test forms (i.e., overall difficulty level is dependent on the stations).

Because the raters (or panelists) were nested within each subpanel, the *rater* effect cannot be interpreted without the associated nested component of *panel*. The rater-related effects were the next group of facet effects that were examined: *rater: panel* accounting for 0.6 per cent of total variance; *station* x (*rater: panel*) explaining 1.6 per cent of total variance and; *person* x (*rater: panel*), accounting for 0.2 per cent of total rating variance. These results indicate that about 2.5 per cent of the total rating variance was due to the *rater* nested within the *panel*. In other words, the pass score was very similar across raters.

Next, we examined the panel-related effects: *panel*, the *person* x *panel* and *station* x *panel* effects accounted for essentially no rating variance. These results indicate that there was a negligible amount of variance due to the two subpanels and that the pass score was nearly identical, irrespective of subpanel.

The G-coefficient and D-coefficient for this model [*person* x *station* x (*rater: subpanel*)] were 0.81 and 0.79, respectively, which indicates that the ratings provided for this standard-setting exercise would generalize quite well if a different set of candidates, raters or subpanels were to be used. These results would generalize less well if a different set of stations were to be used since most of the variance is associated with person x station, which indicates that the pass score established for this exam is dependent on the set of stations used to set the standard and would necessitate that test score linking be implemented to ensure comparability of this standard across test forms (Kolen

& Brennan, 2004). Relating to this point, please note that test score linking is conducted in subsequent sessions to the October 2018 administration to ensure comparability of the pass score for the MCCQE Part II examination.

Table 3: Results of generalizability theory variance component estimates

Facet	df	SS	EMS	EVC	% Variance
Person	49	2089.16	42.64	0.17	**22.6%**
Station	9	419.30	46.59	0.04	**5.0%**
Panel	1	1.06	1.06	0.00	**0.0%**
Person x station	441	3635.43	8.24	0.41	**53.3%**
Person x panel	49	5.85	0.12	0.00	**0.0%**
Station x panel	9	6.61	0.73	0.00	**0.0%**
Person x station x panel	441	54.52	0.12	0.00	**0.0%**
Rater: panel	18	58.40	3.24	0.00	**0.6%**
Person x (rater: panel)	882	128.87	0.15	0.00	**0.2%**
Station x (rater: panel)	162	120.89	0.75	0.01	**1.6%**

df = degrees of freedom SS = sums of squares
EMS = Expected mean squares
EVC = Estimated variance components
% Variance = Percentage of total variance

3.3. IMPACT DATA – PASS RATES

In Table 4, the pass rate for the Initial Round and the Final Round are shown for the CMG-CPG first-time candidates, first-time test takers and all candidates (or total) for the MCCQE Part II October 2018 test form. The overall pass rate is slightly lower for the Final Round as compared to the Initial Round as the pass score increased between the Initial Round and the Final Round.

Table 4: Pass rates by round and candidate cohort for October 2018 exam[1]

Candidate cohort	Round 1	Round 2
CMG-CPG first-time test takers	91.0%	90.2%
First-time test takers	84.0%	83.1%
Total (all test takers)	78.8%	77.9%

3.4. HOFSTEE RESULTS

We computed the Hofstee results for each panel as a function of round (Initial Round and Final Round; see Table 5). Initial Round and the Final Round ratings were similar across subpanels.

[1] The pass rate presented in Table 4 is based on unrounded percentage scores and excluded 151 special case candidates. The pass rate for the entire cohort will be reported in the October 2018 MCCQE Part II technical report after it is transformed to the reported score scale and special cases are decided by the CEC.

There were slight differences between the Initial Round and the Final Round results within each subpanel. All the ranges provided by the panelists fall within the Borderline Group pass scores shown in Table 2. This indicates that the panelists' "gut" estimates were in line with the results based on the Borderline Group method.

Table 5: Summary of Hofstee results by round and panel

Round	Statistic	Subpanel 1	Subpanel 2	Across Panels
Initial round	Percent min.	49.5%	44.2%	**46.9%**
	Percent max.	72.5%	69.5%	**71.0%**
	Failure min.	8.4%	8.7%	**8.6%**
	Failure max.	24.0%	24.0%	**24.0%**
Final round	Percent min.	47.5%	41.5%	**44.5%**
	Percent max.	69.0%	71.0%	**70.0%**
	Failure min.	6.4%	7.6%	**7.0%**
	Failure max.	21.1%	22.8%	**22.0%**

3.5. SUMMARY OF EVALUATION SURVEY FINDINGS

Finally, we divided the evaluation survey into sections that largely reflect major activities that occurred over the three-day meeting. See Appendix E for a full summary of the survey across all panelists and by subpanel with each survey question and results presented. Overall findings of the survey indicate that:

1. All panelists were *very clear*, *clear*, or *somewhat clear* on the description of the *just qualified/borderline pass* candidate. 95 per cent (n = 19) indicated they were *very clear* or *clear*.
2. 100 percent (n = 20) of the panelists indicated that they benefitted from the discussion of the *just qualified/borderline pass* candidate early in the meeting. 90 per cent (n = 18) of the panelists thought the time spent on the description was *about right*; while 10 per cent (n = 2) would have liked *more time* devoted to this activity.
3. 95 per cent (n = 19) of panelists felt that the length of time for the training session was *about right*. Similarly, 90 per cent (n = 18) indicated that the clarity of scoring procedures was *excellent* or *very good*. 85 per cent (n = 17) of the panelists rated the training of the process for setting the pass score as *excellent* or *very good*, while 15 per cent (n = 3) of the panelists rated the training as *good*.
4. Panelists were asked what factors influenced their ratings. All of the factors we considered important were indicated by some or many of the panelists: the description of the *just qualified* candidate (n = 18), panelist discussion (n = 18), experience and knowledge of the field (n = 17), and knowledge and skills measured by the stations (n = 17). Least frequently cited by the panelists were the station statistics (n = 9), statistical impact data before the final round (n = 3).

5. With regard to allotted time, 85 per cent (n = 17) of the panelists judged the time as *about right* for rating the candidate score sheets; the remaining 15 per cent (n = 3) felt *too much* time was allowed. No panelist noted feeling "rushed" in completing their ratings.

6. 65 per cent (n = 13) of the panelists were *very comfortable* with the individual panel discussions while 25 per cent (n = 5) reported being *somewhat comfortable* participating in the discussions. One panelist (n = 1) reported being *unsure*, and a different panelist (n = 1) reported being *somewhat uncomfortable*.

7. On the question of the level of confidence that the impact data and final discussion had on arriving at a defensible pass score, 85 per cent (n = 17) of panelists reported being *very confident* or *confident* while 15 per cent (n = 3) reported being *somewhat confident*.

8. Finally, with respect to the most important question, i.e., "What level of confidence do you have in the final recommended pass score for the MCCQE Part II?" 90 per cent (n = 18) of the panelists indicated they were *very confident* (60 per cent; n = 12), or *confident* (30 per cent; n = 6). Two panelists indicated being *somewhat confident* whereas no panelist indicated that he/she was *not at all confident*.

4. Conclusions

Several important aspects of this standard-setting exercise highlight our confidence in the resulting pass score presented to the CEC for their consideration. First, the results of the pass score across panels were very similar. This indicates that several of the factors in the planning and execution of the standard-setting exercise achieved the desired outcome, which was a fair, balanced and valid process for arriving at the recommended pass score. These factors include the selection and assignment of panelists to each subpanel, ensuring common understanding of the performance level definitions provided to the panelists, the training of panelists, and similar processes used to collect panelists ratings. The similar pass scores by subpanel indicate that the pass score can generalize across at least two matched subpanels.

The generalizability results provided additional validation of the result of this standard-setting exercise. The effects of individual panelists were very small, and the effect of subpanel was virtually nil. These results imply that the two subpanels performed in very similar manners, and even more importantly that individual panelists seemed to have a similar perception of an *acceptable/good*, or *just qualified/borderline pass*, and *unacceptable/poor* candidate. The generalizability analyses evaluated whether the candidate score sheets were rated in the same way for the *acceptable/good* and *unacceptable/poor* categories, in addition to judgments for the candidate score sheets that were classified as *just qualified/borderline* pass. The similar pass scores by subpanel indicate high similarity in judgments of the just qualified/borderline pass candidate score sheets, but the G-analyses evaluated all ratings from 1 to 3 for all 50 candidates.

The Hofstee results provided a “gut” check that the pass score established by subpanel and across panels was within acceptable ranges, based on an overall holistic impression. The Hofstee results for both the Initial Round and Final Round provided boundaries that were in line with the panelists’ ratings for the Borderline Group method, as well as resulting pass rates that would ensue based on the October 2018 MCCQE Part II form.

Finally, the results of the survey conducted at the end of the three-day standard-setting exercise were quite positive, indicating that the experience from the panelists’ point of view was *excellent* and that we achieved our intended goals of preparing the panelists appropriately. Ultimately, and most importantly, panelists were *very confident* in the recommended pass score. These results are similar to those found with other standard-setting exercises, including our MCCQE Part I exam, and NAC exam. Ultimately, the survey results provide additional validation evidence in support of the recommended pass score proposed to the CEC.

In summary, the similarity of the pass scores by panel, generalizability results, Hofstee results, impact data being similar to past administrations, and survey results all provide evidence that this standard-setting exercise was validated appropriately. The panel-based standard-setting exercise was a thorough and rigorous process in establishing a pass score and met best practice standards and procedures.

We presented to the CEC the information in this report and impact information for applying this new pass score to the October 2018 candidate results. Using the October 2018 results of all[2] MCCQE Part II candidates, we established the new scale to have a mean of 150 and a standard deviation of 20. On this new scale, the pass score that was recommended from the standard-setting panel and approved by the CEC is 138. This pass score will remain in place for subsequent MCCQE Part II administrations.

[2] Two candidates that did not complete at least nine stations were not included in establishing the new reported score scale.

References

Boulet J., De Champlain, A. F. & McKinley, D. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher, 25,* 245-9.

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., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications Inc.

De Champlain, A. F. (2013). Standard setting methods in medical education. In T. Swanwick (Ed.). *Understanding Medical Education: Evidence, Theory and Practice*. (pp. 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.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer Science + Business Media, LLC.

Norman, G., Bordage, G., Page, G., & Keane, D. (2006). How specific is case specificity? *Medical Education, 40 (7)*, 618-23.

Shavelson, R. J., Webb, N. M. (1991). *Generalizability Theory: A Primer*. Sage Thousand Oaks, CA: Publications Inc.

APPENDIX A:
Invitation letter and demographic sheet

Good afternoon,

To establish a new pass score for Medical Council of Canada Qualifying Examination (MCCQE) Parts I and II, the governing bodies of the Medical Council of Canada (MCC) will launch a standard-setting exercise for each examination. To begin this process, the Psychometrics and Assessment Services (PAS) directorate at the MCC is soliciting participation for two panels to recommend pass scores for the MCCQE Parts I and II. It is expected the final pass score will be applied beginning with the spring 2018 administration for the MCCQE Part I and the fall 2018 administration for the MCCQE Part II.

We hope that you will consider participating in one of our panels, as your clinical expertise and past experience are vital to the success of this standard-setting exercise. We are issuing notice to solicit participants from which we will assemble the panels to help ensure the medical experts and clinical practice contexts across Canada are well represented. Prospective panel members will be selected for only one of the two standard-setting exercises.

Individuals who contributed to test development or scoring processes for the MCCQE Part I and/or the MCCQE Part II in the past few years will not be selected as a panelist for the exam to which they had contributed; the validity of the pass score lies with a separation of test development and scoring processes from standard-setting processes.

Selected panelists will participate in the standard-setting exercise on June 18-19, 2018 for the MCCQE Part I or December 10-12, 2018 for the MCCQE Part II. These exercises will take place at the MCC's offices in Ottawa. Panelists will be guided through a set of procedures to evaluate examination materials to set the pass score. In addition to reasonable travel expenses (see the MCC's travel policy), an honorarium of $600 per day will be provided.

We hope that you will be interested in participating. Should you be, we ask that you complete the demographic information survey by September 15, 2017, and tentatively reserve the standard-setting dates in your calendar. Your participation will be confirmed by October 20, 2017. Should you have any questions, please contact us at research@mcc.ca.

Thank you very much for your interest and support in achieving the highest level of medical care for Canadians through excellence in evaluation of physicians.

Sincerely,
Director, Associate Director

.....

Psychometrics and Assessment Services | Psychométrie et services docimologiques
MEDICAL COUNCIL OF CANADA | LE CONSEIL MÉDICAL DU CANADA
mcc.ca

Demographic information sheet
Medical Council of Canada standard-setting demographics survey

The information requested below is being collected to help the Medical Council of Canada (MCC) select two representative pan-Canadian panels to recommend a passing score on the Medical Council of Canada Qualifying Examination (MCCQE) Part I and Part II. The standard-setting exercises will take place on:

- June 18-19, 2018 (MCCQE Part I)
- December 10-12, 2018 (MCCQE Part II)

Completed surveys must be submitted by September 15, 2017. Should you have any questions, please contact us at research@mcc.ca.

1. Please provide your full name and contact information.

- Full name:
- Email address:
- Phone number:

2. Do you have your Licentiate of the Medical Council of Canada (LMCC)?

- No
- Yes (please provide your LMCC number)

3. Which of the following certifications do you have? Please select all that apply.

- Royal College of Physicians and Surgeons of Canada (RCPSC)
- College of Family Physicians of Canada (CFPC)
- Collège des médecins du Québec (CMQ)
- None of the above

4. Do you have an unrestricted licence to practise?

- No
- Yes (please specify which province/territory):

5. Number of years in practice post-residency:

- 0-2 years
- 3-5 years
- 6-10 years
- 11-20 years
- 21-30 years
- More than 30 years

6. Number of years experience supervising residents:

- 0-2 years
- 3-5 years
- 6-10 years
- 11-20 years
- 21-30 years
- More than 30 years

7. Are you actively supervising students/residents?

- No
- Yes (please specify how often and how many students/residents you typically supervise in a given year):

8. Number of years supervising Canadian medical graduates (CMGs):

- 1-5 years
- 6-10 years
- 11-20 years
- 21-30 years
- More than 30 years
- I have no experience supervising CMGs

9. Have you ever participated in an MCC test committee or content development workshop?

- No
- Yes (please specify the activity and when):

10. Have you ever been a marker for the clinical decision making (CDM) component of the MCCQE Part I or an examiner for the MCCQE Part II? Please select all that apply.

- I have been an MCCQE Part I CDM marker
- I have been an MCCQE Part II examiner
- I have not done either

11. Have you participated in a preparatory course involving the MCCQE Part I or II within the last three years?

- No
- Yes (please specify the activity and when):

12. Where did you complete your postgraduate medical training?

- Canada
- Other (please specify):

13. Region of the country in which you practice:

- Alberta
- British Columbia
- Manitoba
- New Brunswick
- Newfoundland and Labrador
- Northwest Territories
- Nova Scotia
- Nunavut
- Ontario
- Prince Edward Island
- Quebec
- Saskatchewan
- Yukon

14. First language:

- English
- French
- Other (please specify):

15. Primary language of your medical practice:

- English
- French
- Other (please specify):

16. Gender:

- Female
- Male

17. Ethnicity:

- Caucasian
- Indigenous
- Visible minority (please specify):

18. Medical specialty:

- Pediatrics
- Internal medicine
- Psychiatry
- OBGYN

- Surgery
- Family medicine
- Other (please specify):

19. Type of community in which you primarily work:

- Urban
- Rural

20. Type of care setting in which you primarily work:

- Hospital-based setting
- Community-based setting

Individuals who have been involved with the MCCQE Part I in the past are asked to select the MCCQE Part II standard-setting exercise. Similarly, individuals who have been involved with the MCCQE Part II in the past are asked to select the MCCQE Part I standard-setting exercise.

21. I am interested in and fully available to participate in the following standard-setting exercises (please select all that apply):

- MCCQE Part I (June 18-19, 2018 – two days)
- MCCQE Part II (December 10-12, 2018 – three days)

22. Do you have a preference for one standard-setting exercise over the other?

- MCCQE Part I (June 18-19, 2018 – two days)
- MCCQE Part II (December 10-12, 2018 – three days)
- I have no preference

APPENDIX B:
Agenda

DAY 1: Monday, December 10, 2018

TIME	ACTIVITIES	LEAD
07:45	Breakfast	
08:00	Welcome and introductions	Facilitators
08:15	View security video	Facilitators
08:20	Review the agenda/objectives	Facilitators
08:30	Overview of MCCQE Part II	MCC staff
08:55	Overview of standard setting	MCC staff
09:15	Just qualified candidate discussion	MCC staff
10:00	Break	
10:15	Training for long station	Facilitators
11:45	Lunch	
12:45	Station T01 (Initial round)	Facilitators
14:15	Station T02 (Initial round)	Facilitators
15:35	Break	
15:45	Station T04 (Initial round)	Facilitators
17:00	Wrap-up day 1	

DAY 2: Tuesday, December 11, 2018

TIME	ACTIVITIES	LEAD
07:45	Breakfast	
8:00	Station T05 (Initial round)	Facilitators
9:00	Station T06 (Initial round)	Facilitators
10:00	Break	
10:15	Station T07 (Initial round)	Facilitators
11:15	Station T08 (Initial round)	Facilitators
12:15	Lunch	
13:00	Training paired station	Facilitators
14:00	Station C01 (Initial round)	Facilitators
14:45	Break	
15:00	Station C05 (Initial round)	Facilitators
15:45	Station C07 (Initial round)	Facilitators
16:40	Hofstee	Facilitators
16:55	Wrap-up day 2/Overview of day 3	
18:00	Dinner	

DAY 3: Wednesday, December 12, 2018

TIME	ACTIVITIES	LEAD
08:00	Breakfast	
08:15	Present impact data and discussion	Facilitators
09:00	Station T01 (Final round)	Facilitators
09:30	Station T02 (Final round)	Facilitators
10:00	Station T04 (Final round)	Facilitators
10:30	Break	
10:45	Station T05 (Final round)	Facilitators
11:15	Station T06 (Final round)	Facilitators
11:45	Station T07 (Final round)	Facilitators
12:15	Lunch	
13:00	Housekeeping (e.g., expense claims, taxis, etc.)	MCC staff
13:15	Station T08 (Final round)	Facilitators
13:45	Station C01 (Final round)	Facilitators
14:15	Station C05 (Final round)	Facilitators
14:45	Station C07 (Final round)	Facilitators
15:15	Hofstee	Facilitators
15:25	Break	
16:05	Present impact data	Facilitators
16:35	Post exercise survey	Facilitators
16:50	Wrap-up day 3	

APPENDIX C:
Performance level descriptors MCCQE Part II

The candidate's lowest deficiency in any of Information Gathering, Diagnosis and Management or Interpersonal Skills and Professionalism is how they are categorized overall. For example, if a candidate is Acceptable or Good in Information Gathering and in Diagnosis and Management, but they are Unacceptable or Poor in Interpersonal Skills and Professionalism, they are categorized as Unacceptable or Poor.

Unacceptable or poor candidate	Borderline pass or marginally qualified candidate	Acceptable or good candidate
The candidate is <u>not</u> qualified for independent practice. The deficiencies are such that the candidate may put the patient at risk, or they may not ensure the patient's basic needs are met.	**The candidate is qualified for independent practice.** The deficiencies are such that the candidate does not put the patient at risk, and they ensure the patient's basic needs are still met.	**The candidate is qualified for independent practice.** If the candidate has any deficiencies, they are minor, and the candidate does not put the patient at risk in any way. The candidate ensures the patient's basic needs are fully met.
Information gathering		
The candidate usually demonstrates incomplete or disorganized information gathering from the history, physical examination or laboratory data. Even when the candidate gathers sufficient information, their approach is disorganized, and their physical examination technique is poor.	The candidate demonstrates an ability to gather most of the patient's essential information (including laboratory data), but aspects of their history gathering, or physical examination may be disorganized or may lack the expected skill to consistently develop a clear definition of the patient's problem.	The candidate demonstrates an ability to gather information of sufficient breadth and depth to develop a clear definition of the patient's problem through history gathering, a logical physical examination and appropriate investigations.
Diagnosis and management		
Gaps in the candidate's information gathering or their interpretation of information results in an incoherent differential diagnosis or incomplete management plan.	Gaps in the candidate's information gathering or their interpretation of information may affect the breadth and depth of their differential diagnosis and may reduce the completeness of their management plan.	The candidate gathers sufficient information and interprets it accurately, presents it logically and prioritizes it to make an appropriate differential diagnosis. Based on the diagnosis, the candidate consistently provides appropriate management.
Interpersonal skills and professionalism		
The candidate demonstrates little ability to engage with the patient, is not patient-centered and is not sensitive to the patient's needs or the patient's understanding of the information provided. The candidate appears to lack confidence or appears over-confident during their interactions with the patient. The candidate does not exhibit professional behaviour.	While the candidate may gather or provide information from or to the patient or others, their approach may not always be patient-centered. The candidate may not consistently respond to the patient's verbal and non-verbal cues regarding the patient's understanding of the information provided. The candidate exhibits professional behaviour.	The candidate puts the patient at ease, consistently shows respect, demonstrates a patient-centered approach in gathering and providing information and verifies the patient's understanding of any information provided. The candidate exhibits professional behaviour.

APPENDIX D:
Hofstee paper form

Panelist: _____

Subpanel: _____

Round: Initial

1. What is the **highest** percent pass score that would be acceptable, even if every candidate attains that score?
2. What is the **lowest** percent pass score that would be acceptable, even if no candidate attains that score?
3. What is the **maximum** acceptable failure rate?
4. What is the **minimum** acceptable failure rate?

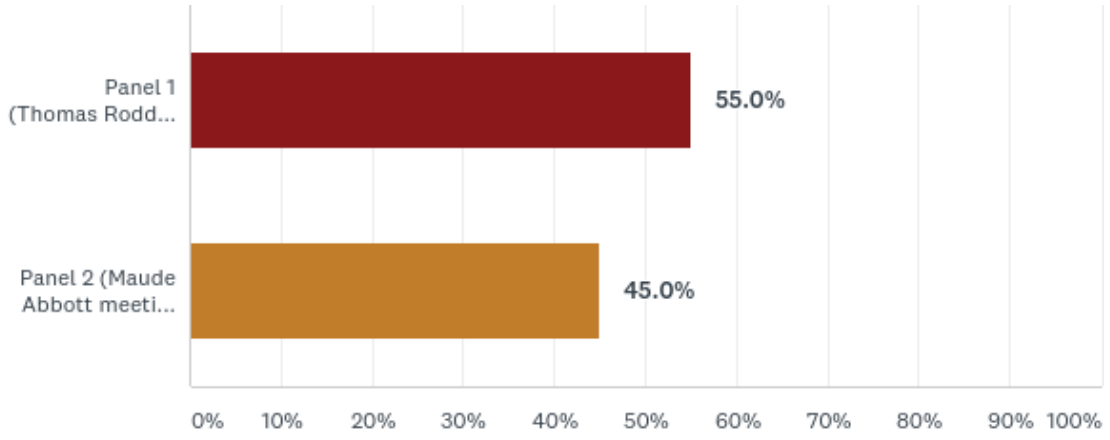
Round: Final

1. What is the **highest** percent pass score that would be acceptable, even if every candidate attains that score?
2. What is the **lowest** percent pass score that would be acceptable, even if no candidate attains that score?
3. What is the **maximum** acceptable failure rate?
4. What is the **minimum** acceptable failure rate?

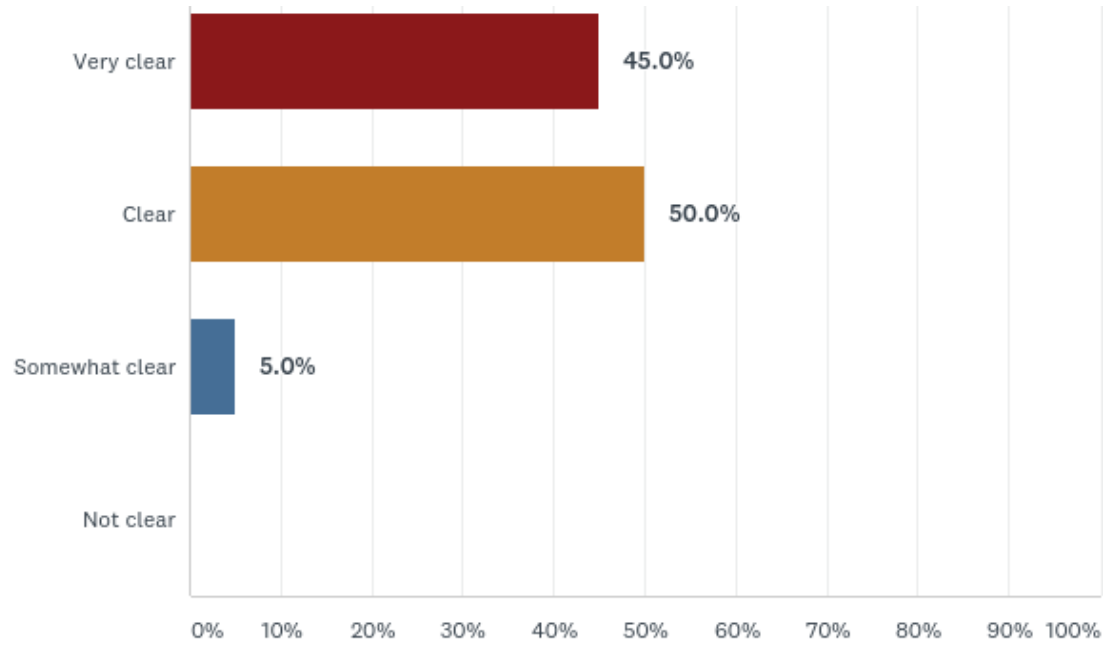
APPENDIX E:
Summary of responses to post-meeting survey

All Panelists

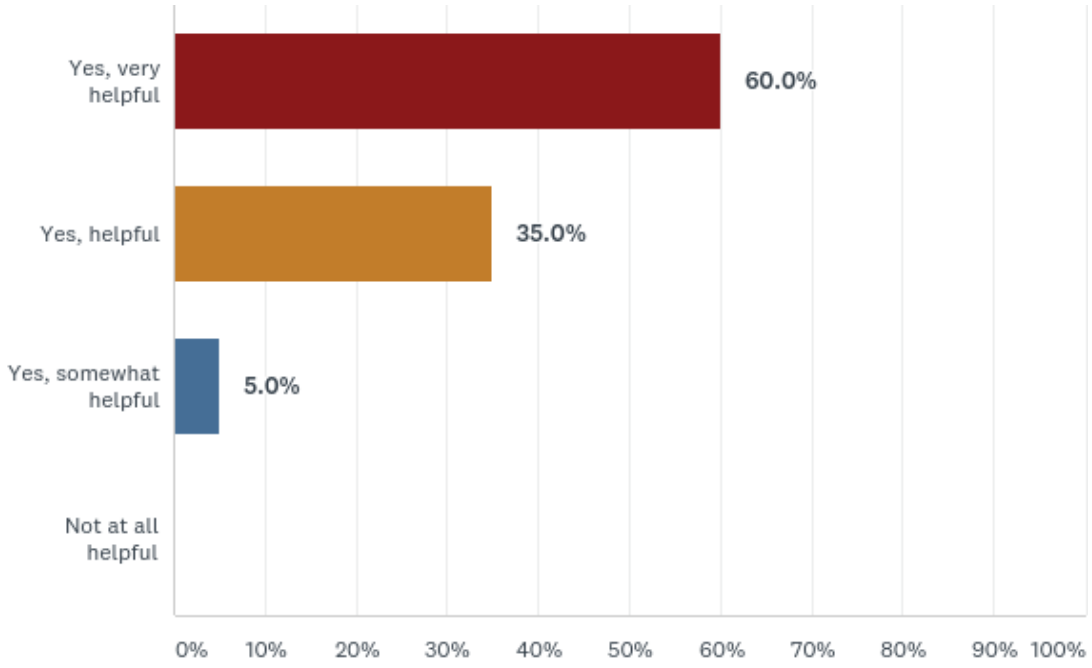
Q1. Which panel did you participate in?



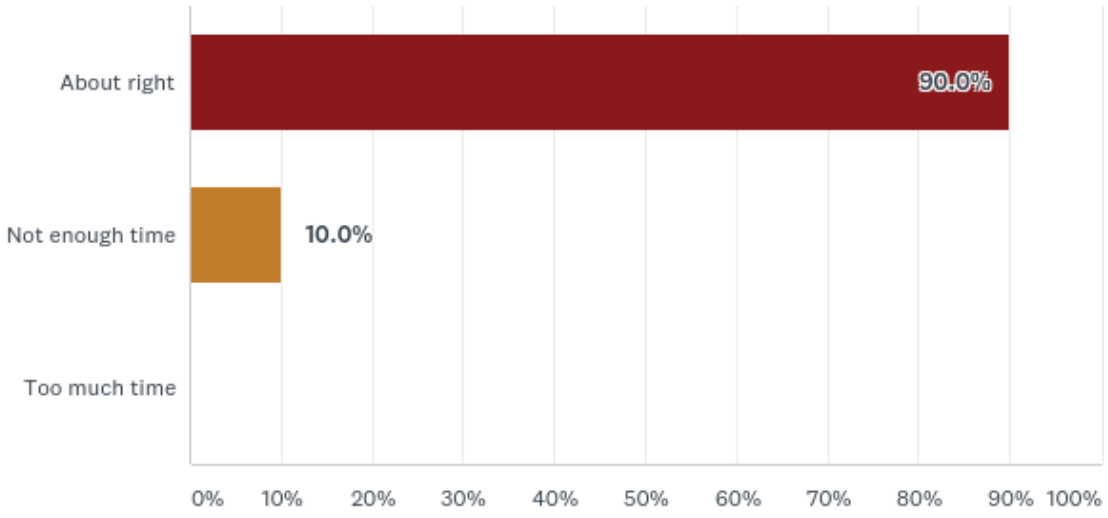
Q2. Following the training on Day 1, how clear was the description of the "Just Qualified" (or "Borderline Pass") candidate on the MCCQE Part II?



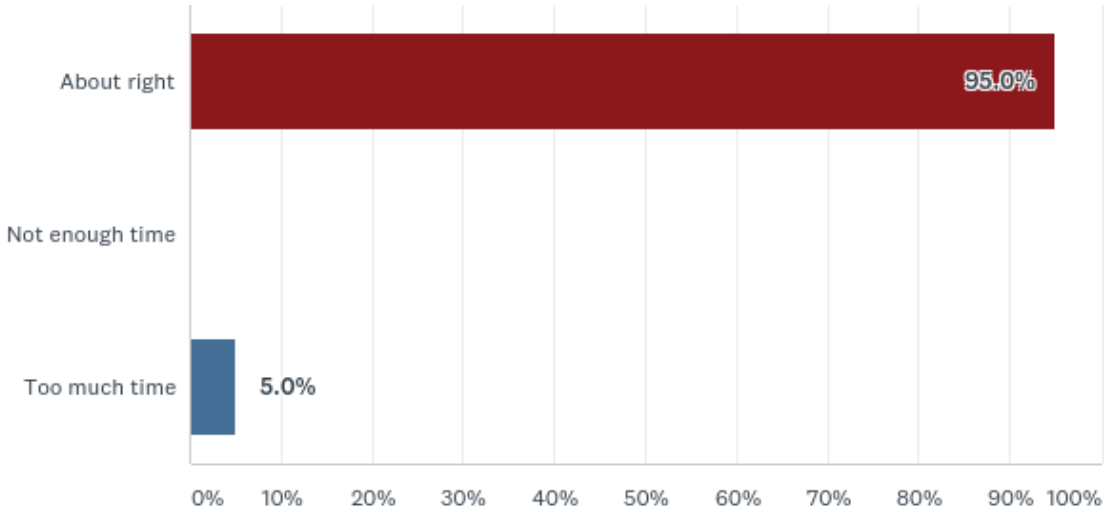
Q3. During the training on Day 1, how helpful was the discussion of the "Just Qualified" (or "Borderline Pass") candidate on the MCCQE Part II?



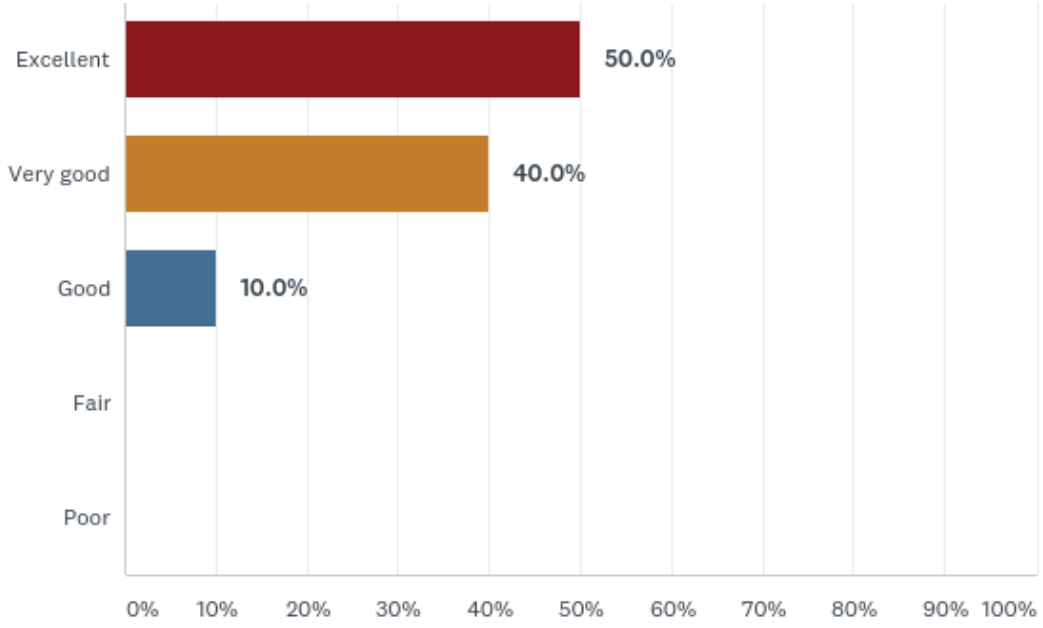
Q4. How would you judge the length of time spent introducing and discussing the description of the "Just Qualified" (or "Borderline Pass") candidate (approximately 45 minutes)?



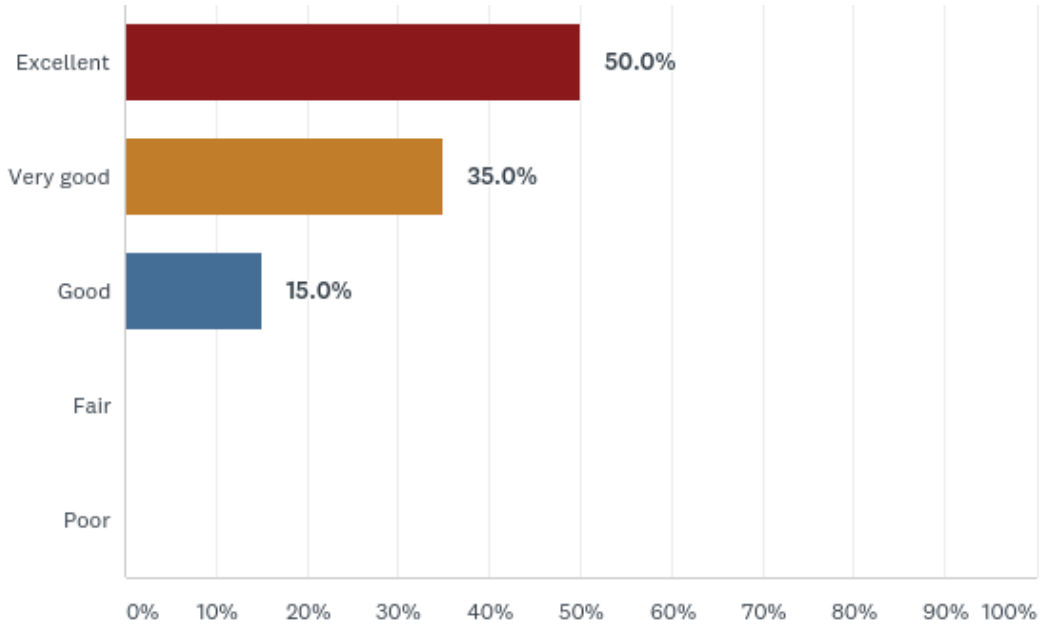
Q5. What is your impression of the length of training time you received for setting a pass score for the MCCQE Part II?



Q6. How clear was the information provided regarding the scoring procedures for the MCCQE Part II?



Q7. What is your overall evaluation of the training provided for setting a pass score for the MCCQE Part II?

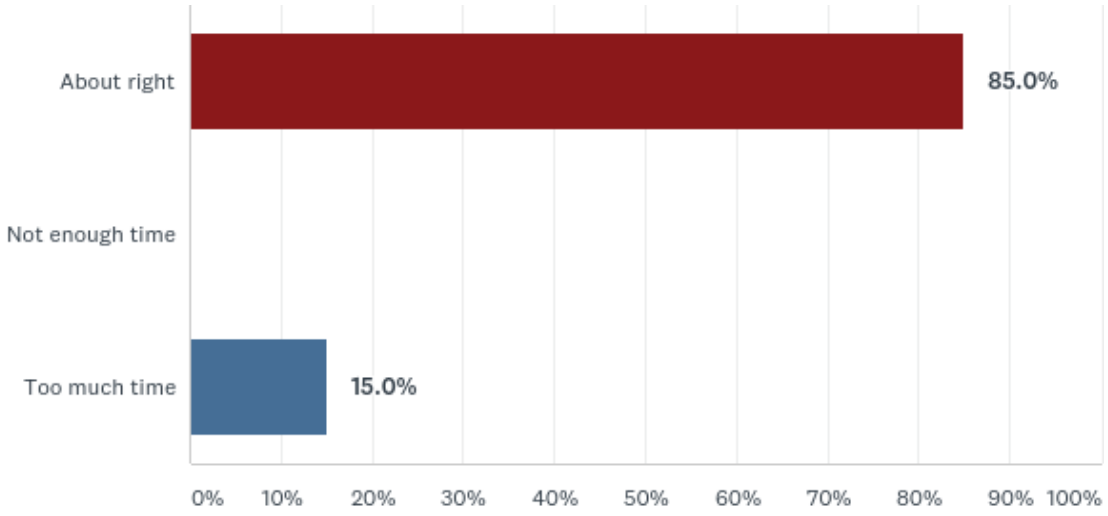


Q8. What factors influenced the ratings you made of the "Just Qualified" (or "Borderline Pass") candidate responses on the MCCQE Part II? Please select all that apply.

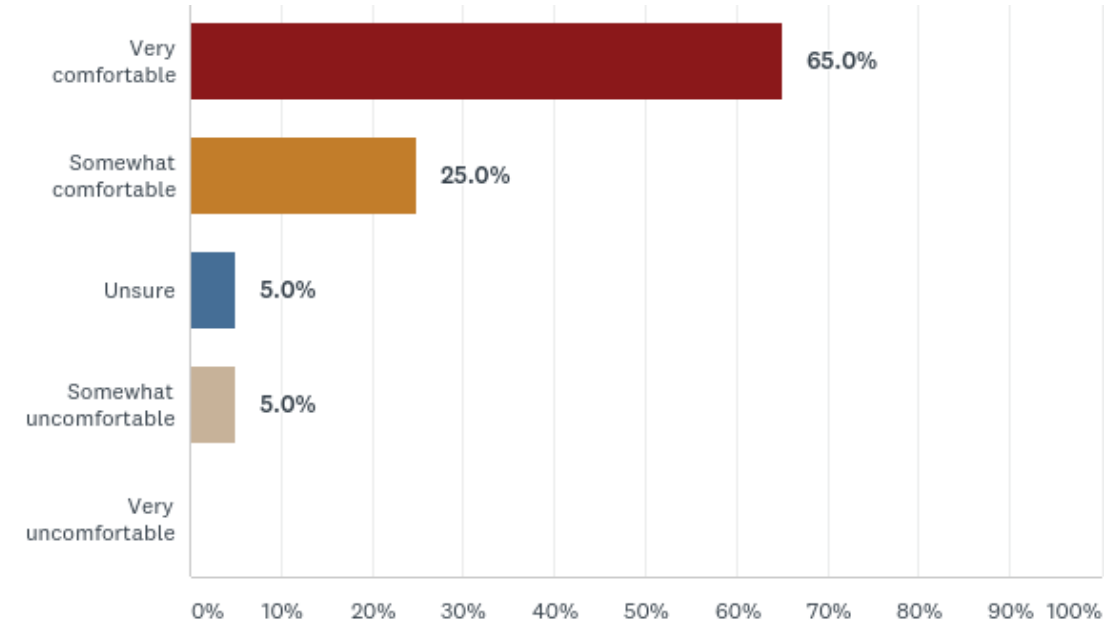
ANSWER CHOICES	RESPONSES	
The description of the "Just Qualified" or "Borderline Pass" candidate	90.0%	18
My perception of the difficulty of the stations or station components	55.0%	11
The scoring of the individual stations or station components	70.0%	14
The station statistics (e.g., candidate station scores)	45.0%	9
The statistical impact data provided before the final round	15.0%	3
Panelist discussions	90.0%	18
My experience in the field	85.0%	17
Knowledge and skills measured by the stations	85.0%	17
Other (please specify):	5.0%	1
Total Respondents: 20		

Other (please specify): *"The completion of critical items on the checklist. Also, the global impressions by the marker were important at the final outcome in these cases."*

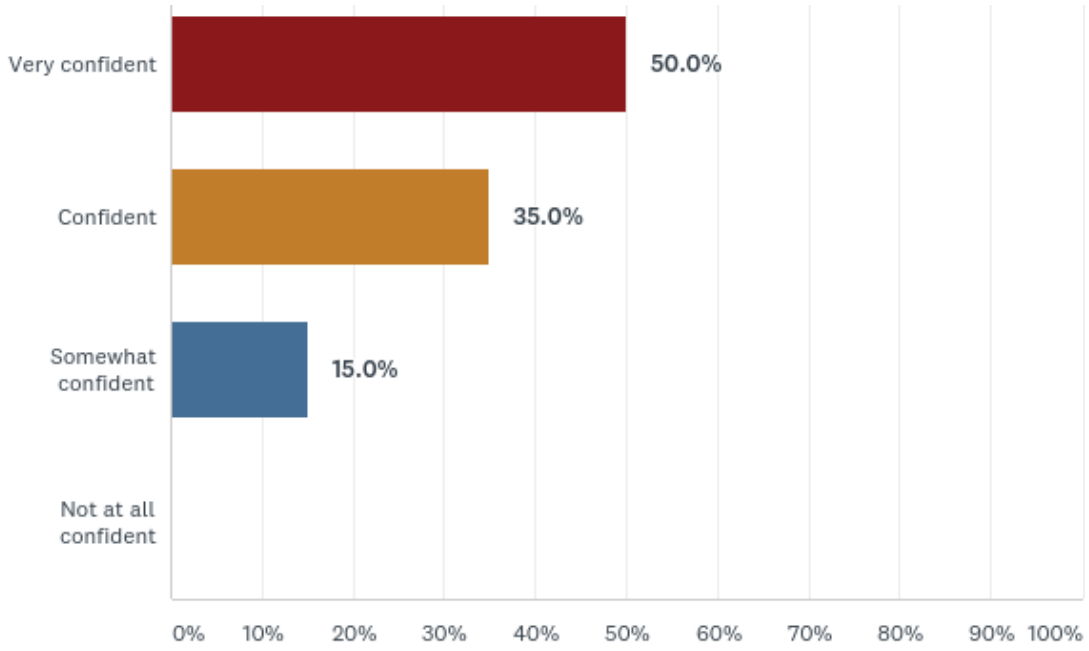
Q9. How would you judge the length of time provided for completing the ratings for each of the stations?



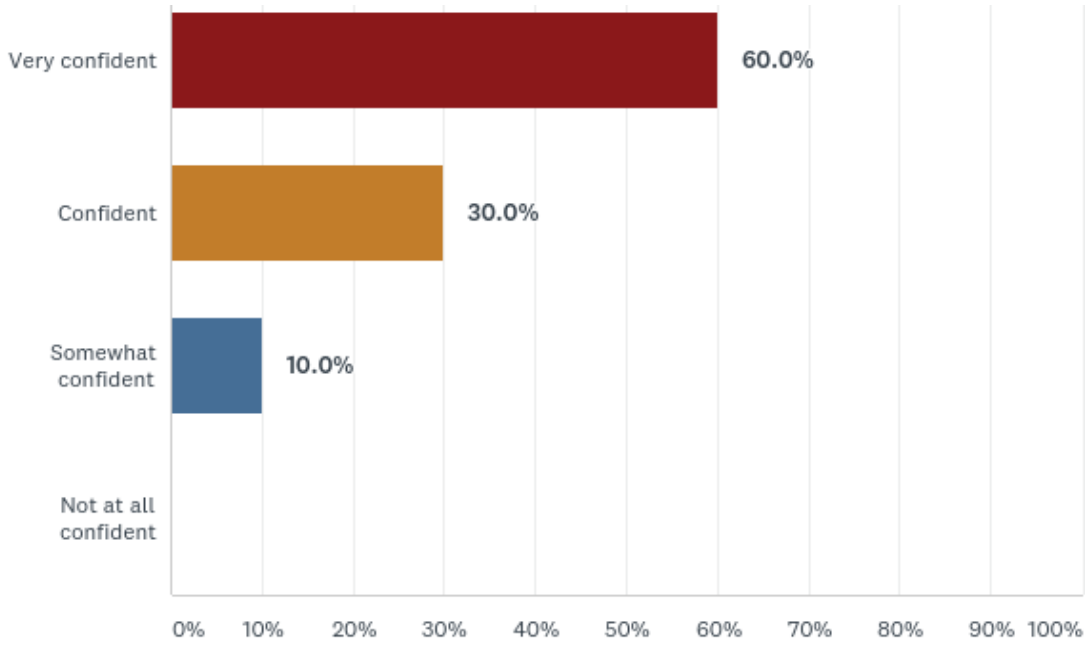
Q10. Overall, how did you feel about participating in group discussions conducted during the ratings process for each station?



Q11. What level of confidence do you have that the impact data and final discussion on the final afternoon helped the panel arrive at a defensible pass score?



Q12. What level of confidence do you have in the final recommended pass score for the MCCQE Part II?



Q13. How could the processes used for setting a pass score for the MCCE Part II have been improved?

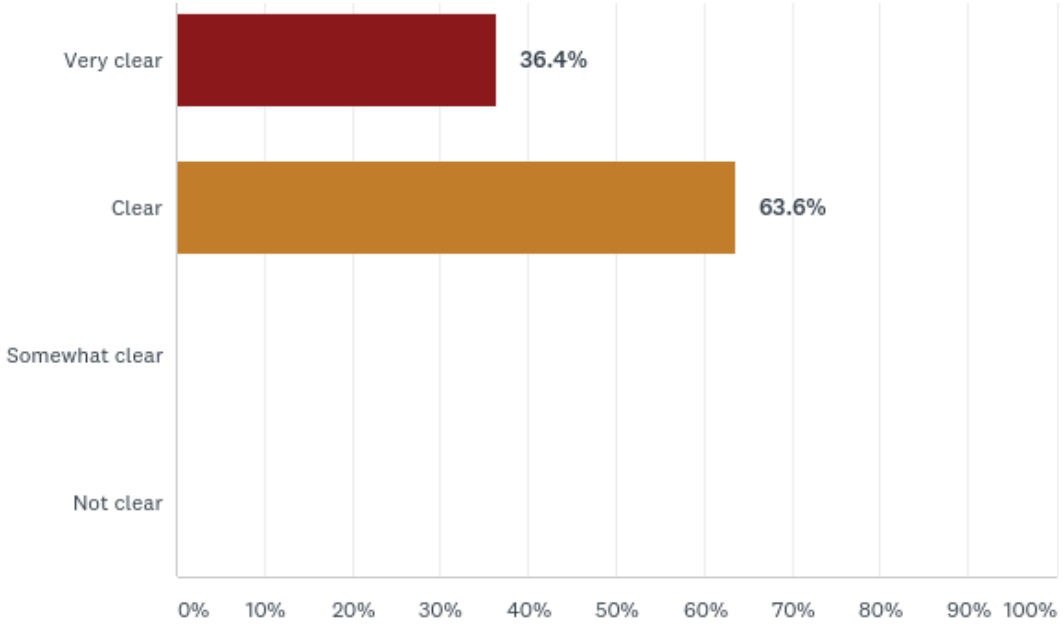
perhaps by reminding us to read the materials that were sent earlier
I believe this arrangement of two rounds of scoring is very appropriate.
no idea
maybe just show one borderline video per station
Wasn't entirely clear to start that could move between 1 and 2 and 3, ie that it was not graduated. A bit more time around that perhaps. Also videos perhaps on more borderline pass and borderline fail rather than really bad or really good. As w proell as more clarity around what attempted vs completed might mean for particular stations. This might mean more specific training even on exam days.
Although discussions about cases (which points were considered more important than others, which points other would fail a candidate for etc.) were interesting and felt helpful, I wonder if there could be bias introduced by this (people being convinced to mark differently than what they would have independently based on their own practice). It would be interesting to see how the cut scores may change if no discussion occurred. Furthermore, I do wonder if there was bias introduced by the ordering of the sheets by score - although this is more time efficient, it also may bias people to being less thoughtful about their marking.
Perhaps editing videos to cut out down-time.
non biased
There is a question about whether borderline candidate videos could be reviewed as well regarding scoring and validation of the mark.
Not applicable
What was done in the second day to tell us not really to look at the scores in the scoring sheets but more on a general impression and feeling about what is more important than the other
possibly have more information given in advance
More time explaining how to review the score sheets and how to use the computer program to input the scores.
See number 14 below - that process actually undermined my confidence in the process as seemed to try and "get people on board" and change their scores.
providing more statistical data for candidates.
establishing essential responses in every station for a pass score
Only show one video vs two videos. Once one knows what the station is about, it should be clear what is borderline and what is not.
computer based evaluation with keeping results of rounds
Shorten time of breaks and lunch to move the process through a bit faster.
Hard to say - we had some great discussions; staff was very clear in their explanations. There appeared to be an excellent cross section of reps. Thanks for allowing me to participate.

Q14. Please provide any additional comments or suggestions about the setting of a pass score for the MCCQE Part II.

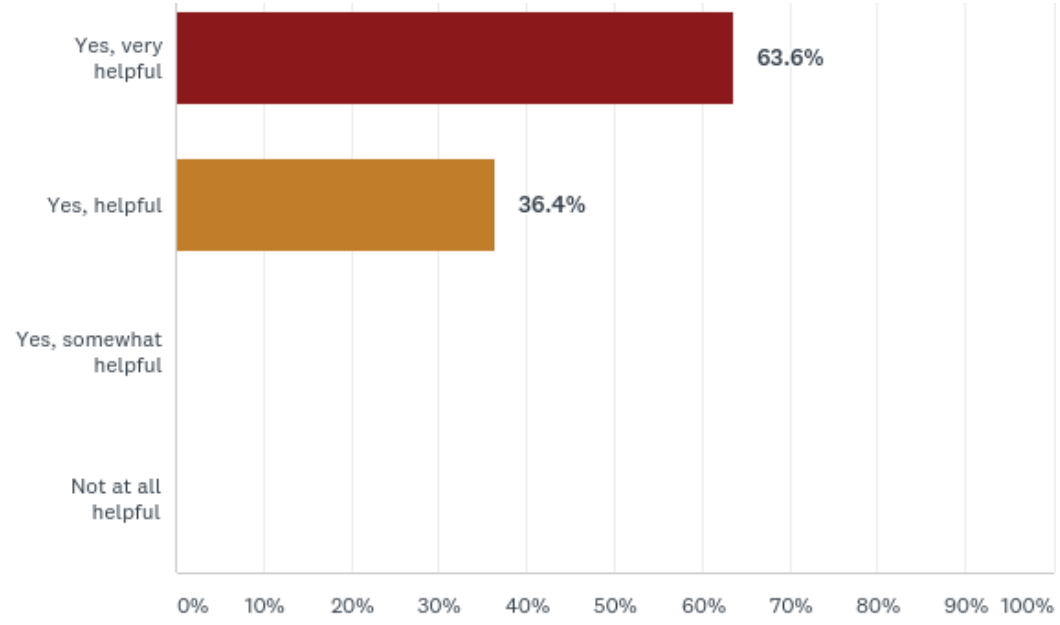
the preparation for this meeting was very thorough
interesting experience. I would do it again someday. hard to make it less mentally straining.
Would be helpful to have the alerts turned off on the computers also. Thanks for putting on a very well organized event. Much appreciation to you all.
The confidence comes from the rigour of the process and the cumulative expertise of the panels. It was eye-opening to appreciate the differences between our observations of examinee performance on the videos and the ratings provided- very subjective. I think it is really important to be consider both items and the global ratings in deliberating an individuals assignment to borderline.
I found the whole process to be very organized and successful
It was a very educational process and I am grateful to have been invited to get the opportunity to participate in this exercise.
This is a very thorough process, and I do expect that more activities like this would continue to be done in order to improve the quality of the MCCQE Part II
Very well organised and welcoming members of the staff. Thank you !
n/a
Appropriately emphasized the importance of this exercise to the participants so that we could understand the impact of the results.
Providing the impact data seemed to be an undue influence on trying to get people to change their scores to that the final appeared to be more homogeneous/consensus. Given the initial impact seemed robust, unclear why needed to re-do.
na

Subpanel 1

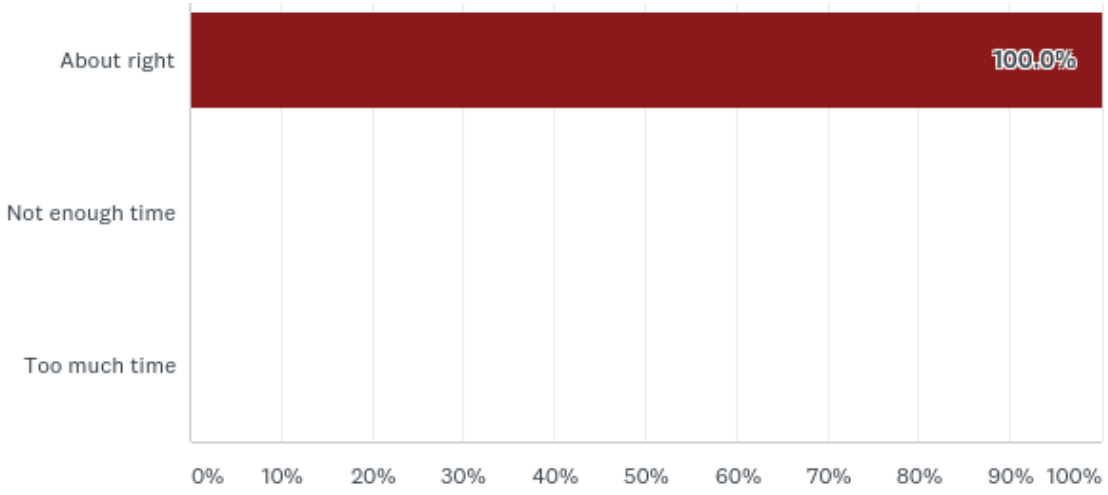
Q1. Following the training on Day 1, how clear was the description of the "Just Qualified" (or "Borderline Pass") candidate on the MCCQE Part II?



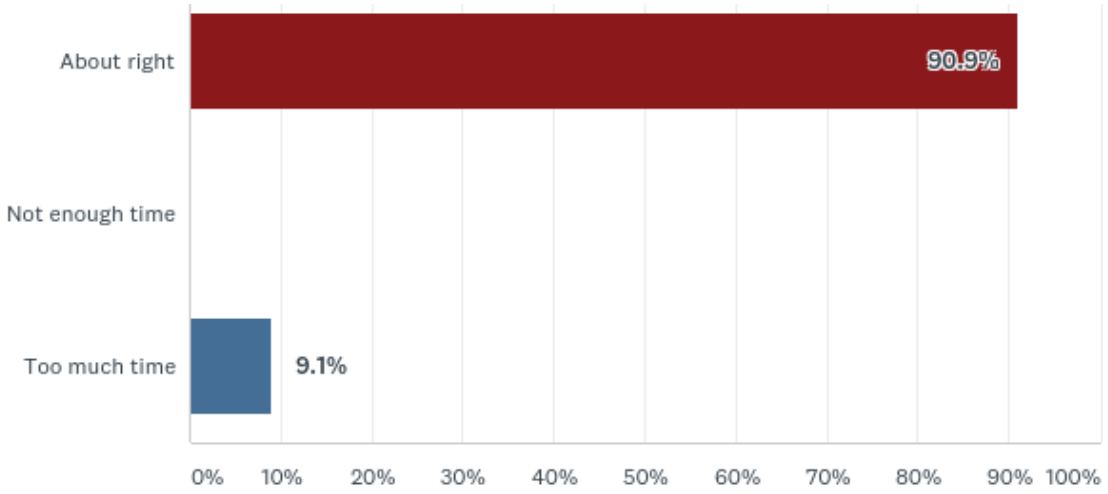
Q2. During the training on Day 1, how helpful was the discussion of the "Just Qualified" (or "Borderline Pass") candidate on the MCCQE Part II?



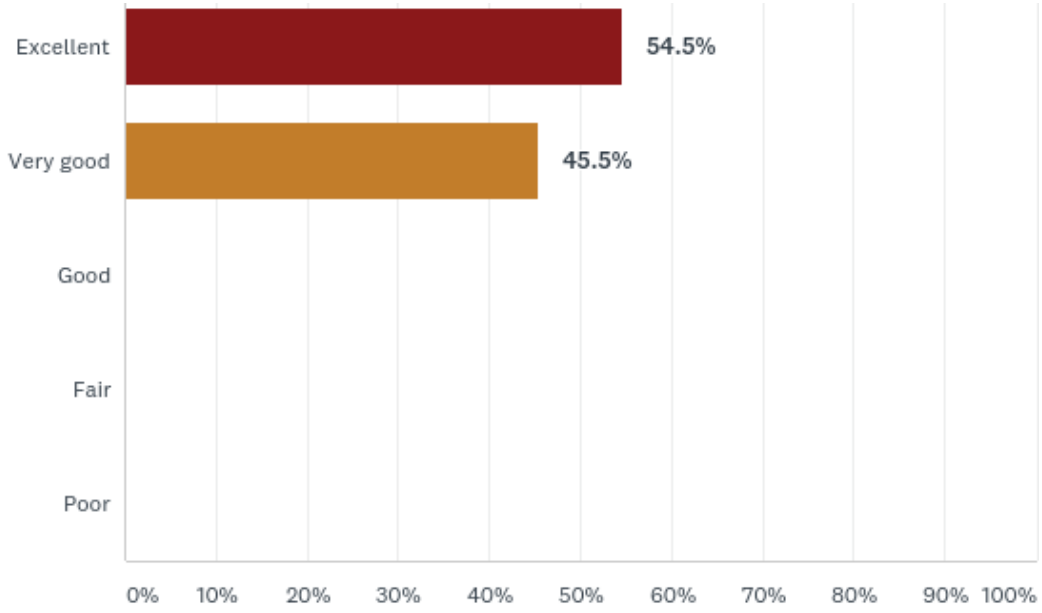
Q3. How would you judge the length of time spent introducing and discussing the description of the "Just Qualified" (or "Borderline Pass") candidate (approximately 45 minutes)?



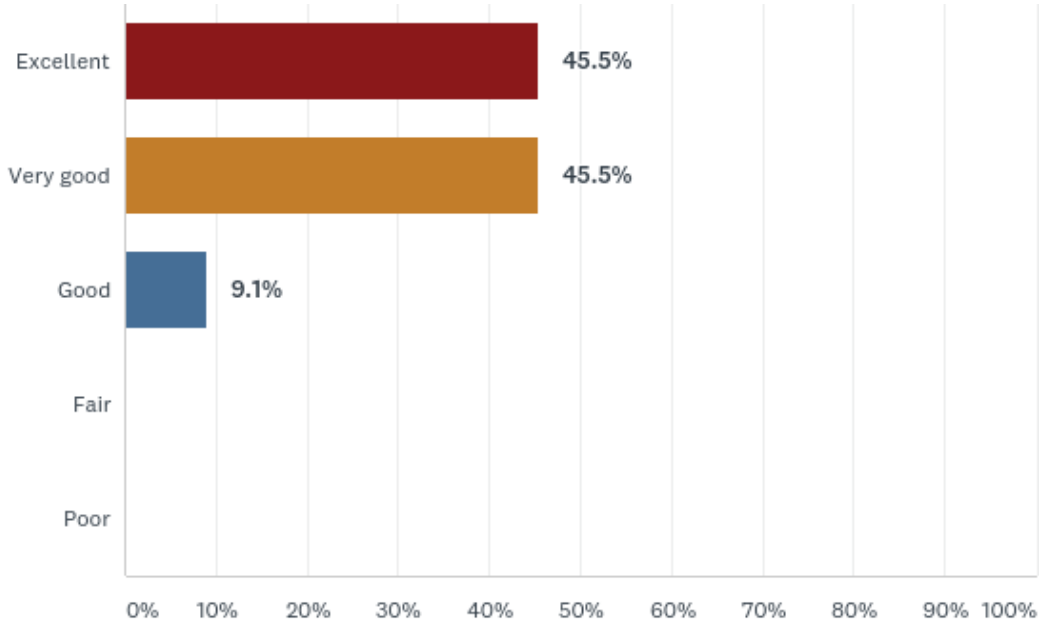
Q4. What is your impression of the length of training time you received for setting a pass score for the MCCQE Part II?



Q5. How clear was the information provided regarding the scoring procedures for the MCCQE Part II?



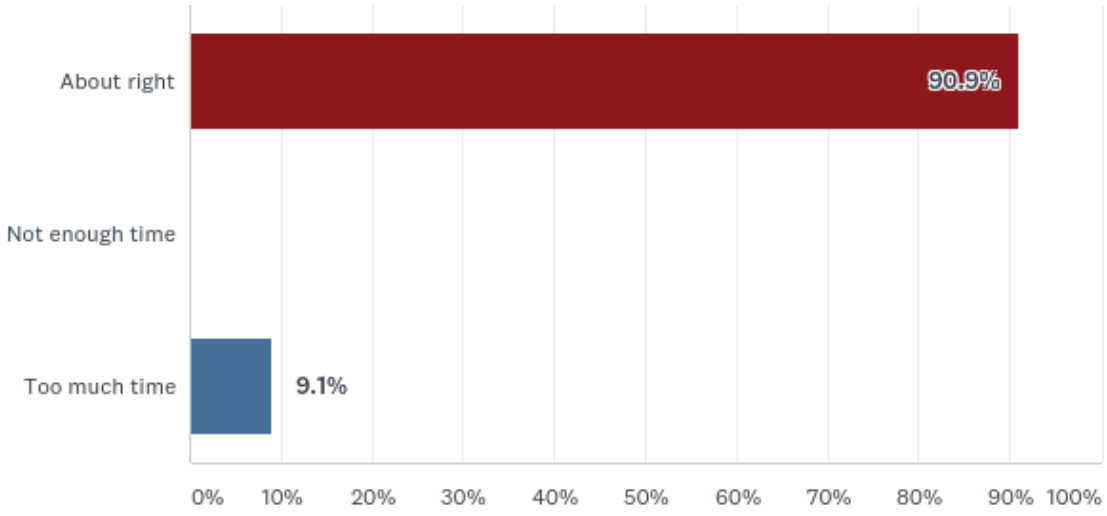
Q6. What is your overall evaluation of the training provided for setting a pass score for the MCCQE Part II?



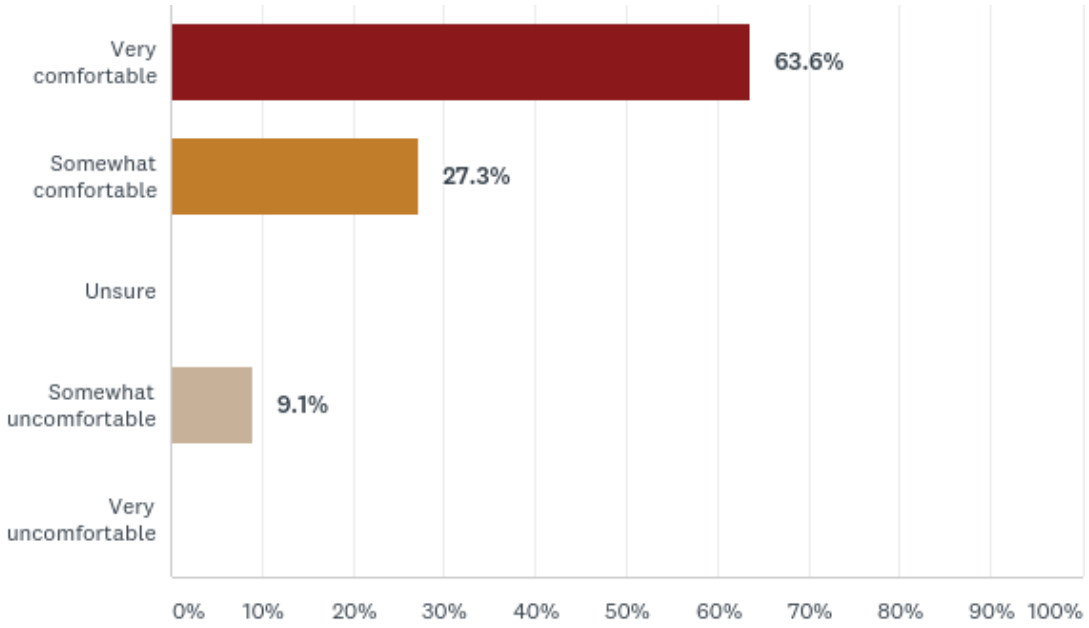
Q7. What factors influenced the ratings you made of the "Just Qualified" (or "Borderline Pass") candidate responses on the MCCQE Part II? Please select all that apply.

ANSWER CHOICES	RESPONSES	
The description of the "Just Qualified" or "Borderline Pass" candidate	90.9%	10
My perception of the difficulty of the stations or station components	45.5%	5
The scoring of the individual stations or station components	63.6%	7
The station statistics (e.g., candidate station scores)	36.4%	4
The statistical impact data provided before the final round	18.2%	2
Panelist discussions	90.9%	10
My experience in the field	90.9%	10
Knowledge and skills measured by the stations	90.9%	10
Other (please specify):	0.0%	0
Total Respondents: 11		

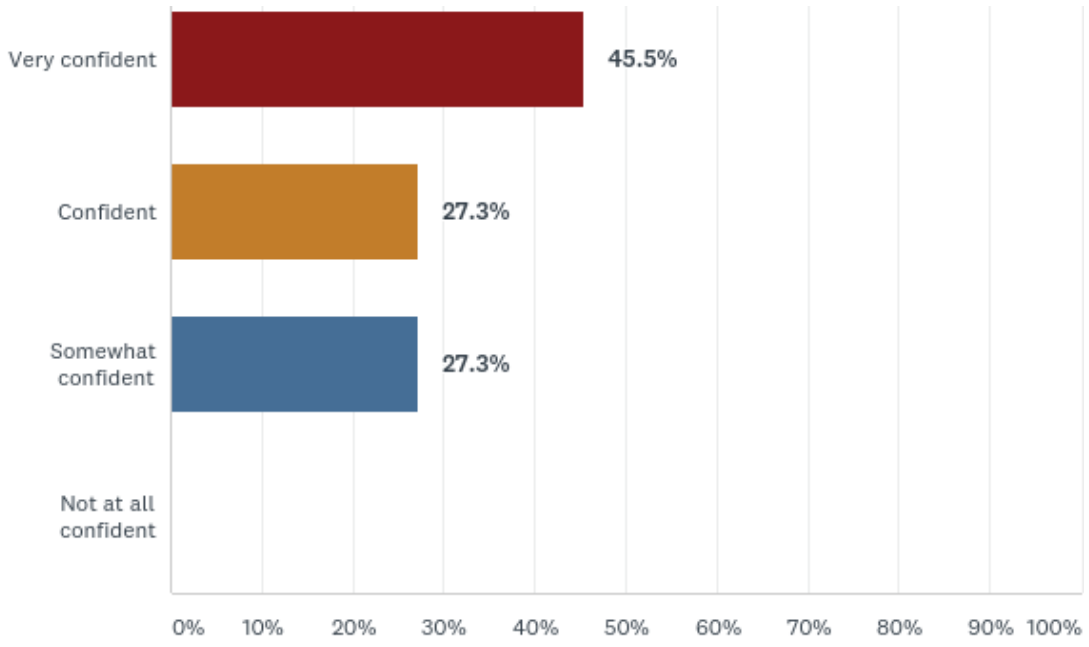
Q8. How would you judge the length of time provided for completing the ratings for each of the stations?



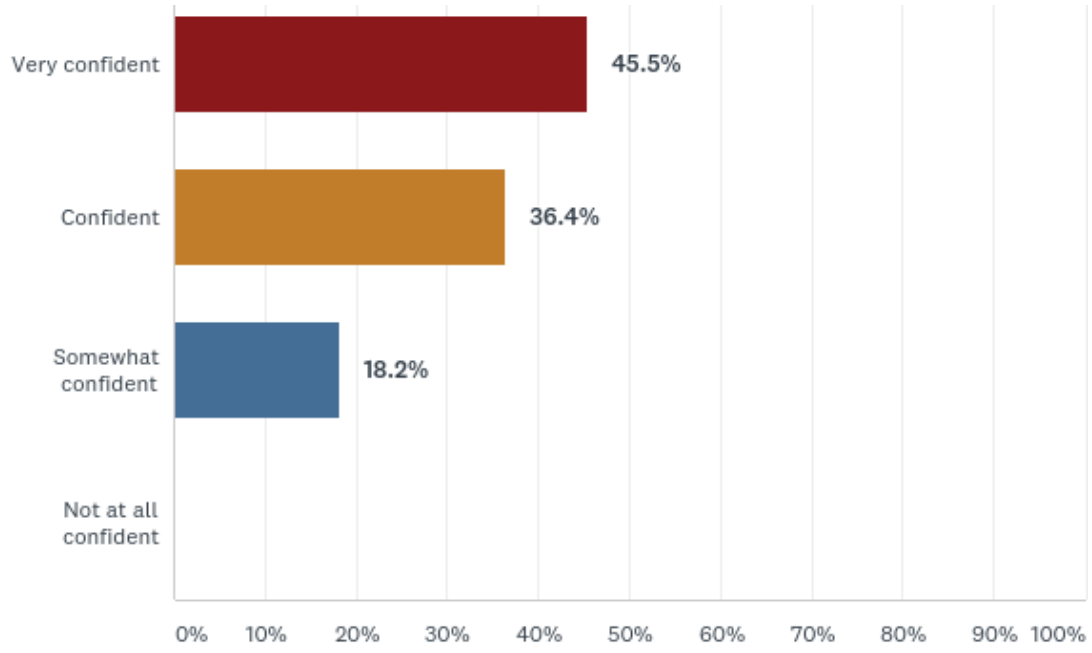
Q9. Overall, how did you feel about participating in group discussions conducted during the ratings process for each station?



Q10. What level of confidence do you have that the impact data and final discussion on the final afternoon helped the panel arrive at a defensible pass score?



Q11. What level of confidence do you have in the final recommended pass score for the MCCQE Part II?



Q12. How could the processes used for setting a pass score for the MCCQE Part II have been improved?

Wasn't entirely clear to start that could move between 1 and 2 and 3, ie that it was not graduated. A bit more time around that perhaps. Also videos perhaps on more borderline pass and borderline fail rather than really bad or really good. As well as more clarity around what attempted vs completed might mean for particular stations. This might mean more specific training even on exam days.
Although discussions about cases (which points were considered more important than others, which points other would fail a candidate for etc.) were interesting and felt helpful, I wonder if there could be bias introduced by this (people being convinced to mark differently than what they would have independently based on their own practice). It would be interesting to see how the cut scores may change if no discussion occurred. Furthermore, I do wonder if there was bias introduced by the ordering of the sheets by score - although this is more time efficient, it also may bias people to being less thoughtful about their marking.
non biased
There is a question about whether borderline candidate videos could be reviewed as well regarding scoring and validation of the mark.
Not applicable

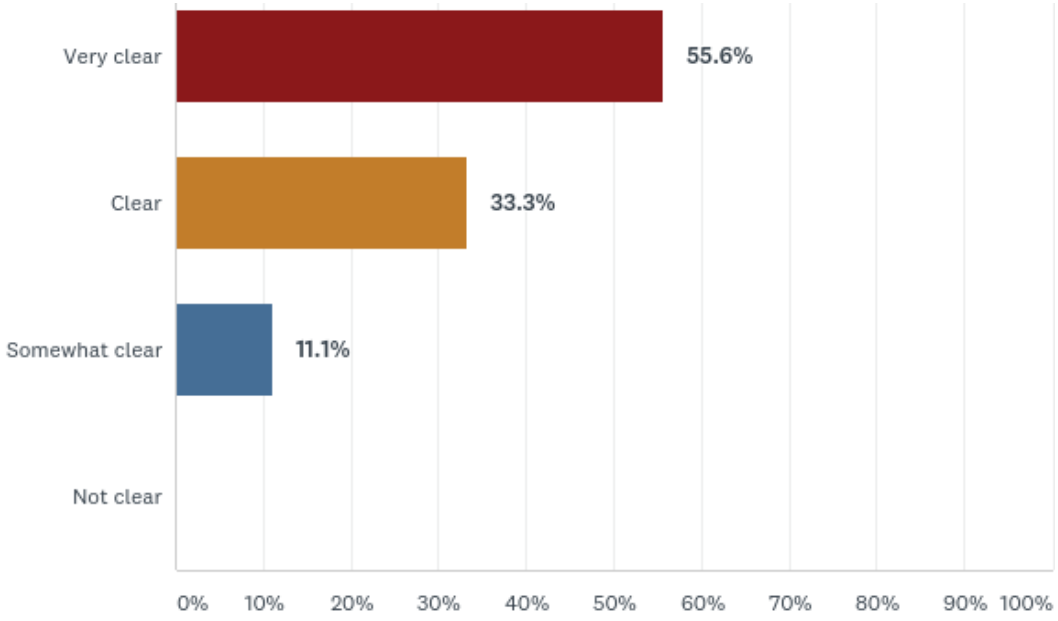
More time explaining how to review the score sheets and how to use the computer program to input the scores.
See number 14 below - that process actually undermined my confidence in the process as seemed to try and "get people on board" and change their scores.
providing more statistical data for candidates.
establishing essential responses in every station for a pass score
Only show one video vs two videos. Once one knows what the station is about, it should be clear what is borderline and what is not.
computer based evaluation with keeping results of rounds
Hard to say - we had some great discussions; staff was very clear in their explanations. There appeared to be an excellent cross section of reps. Thanks for allowing me to participate.

Q13. Please provide any additional comments or suggestions about the setting of a pass score for the MCCQE Part II.

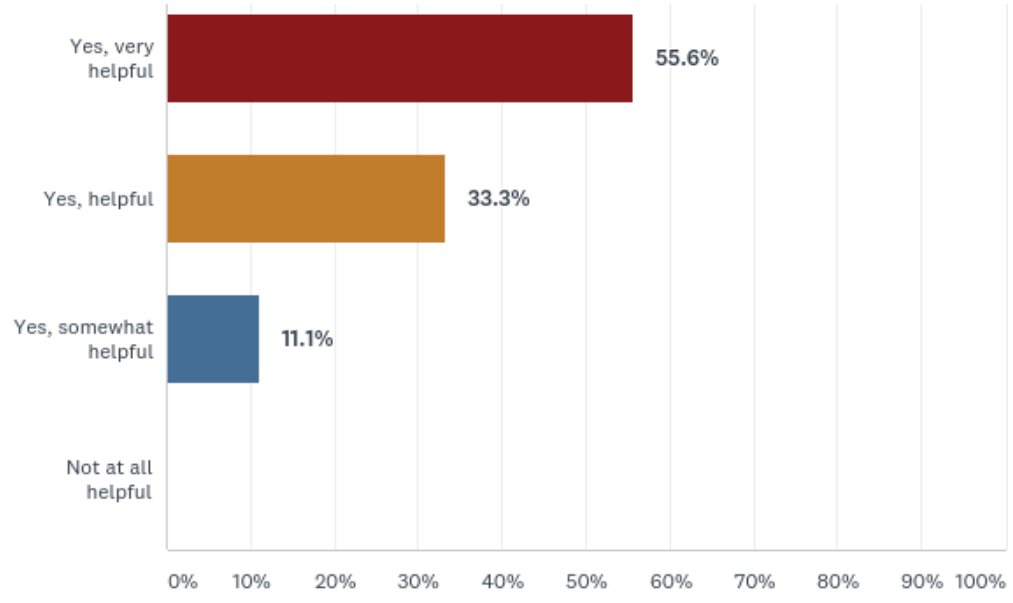
Would be helpful to have the alerts turned off on the computers also. Thanks for putting on a very well organized event. Much appreciation to you all.
I found the whole process to be very organized and successful
This is a very thorough process, and I do expect that more activities like this would continue to be done in order to improve the quality of the MCCQE Part II
Appropriately emphasized the importance of this exercise to the participants so that we could understand the impact of the results.
Providing the impact data seemed to be an undue influence on trying to get people to change their scores to that the final appeared to be more homogeneous/consensus. Given the initial impact seemed robust, unclear why needed to re-do.
na
It was an eye opening experience. I appreciate more the effort that goes into designing an exam like the MCCQE.
See above comment box - little to add as I am not an expert in pyscho-metrics and score setting; but it was very clear that the staff were.

Subpanel 2

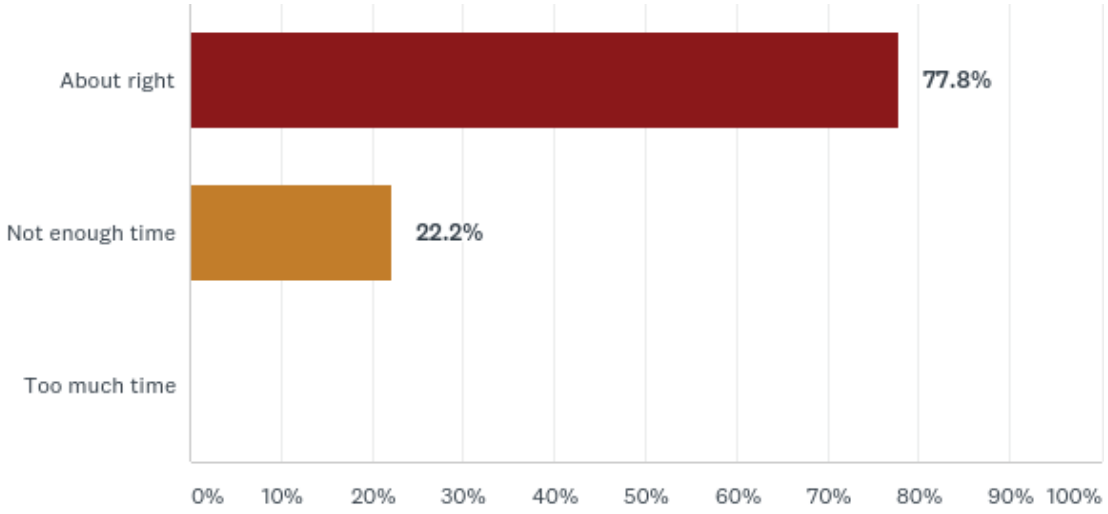
Q1. Following the training on Day 1, how clear was the description of the "Just Qualified" (or "Borderline Pass") candidate on the MCCQE Part II?



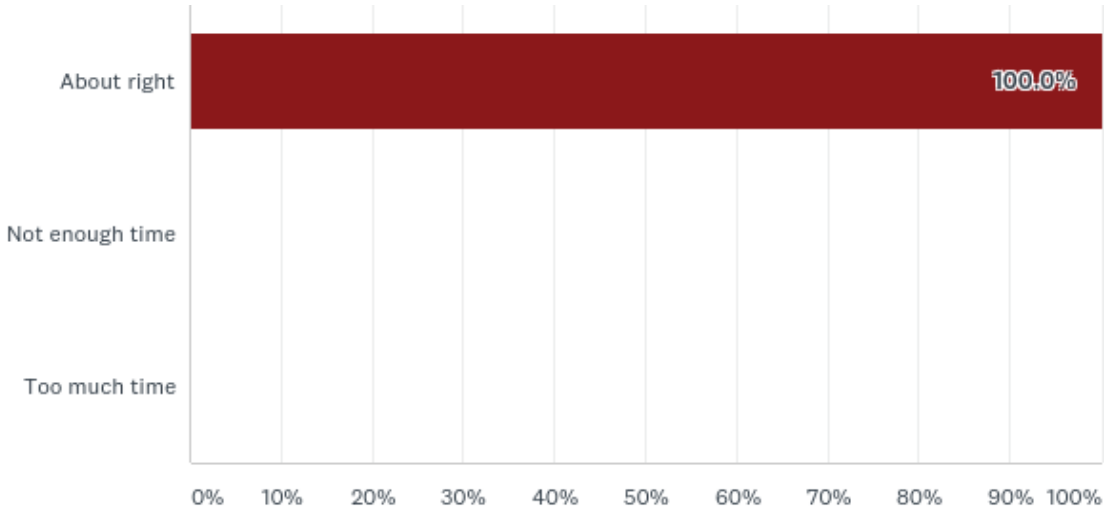
Q2. During the training on Day 1, how helpful was the discussion of the "Just Qualified" (or "Borderline Pass") candidate on the MCCQE Part II?



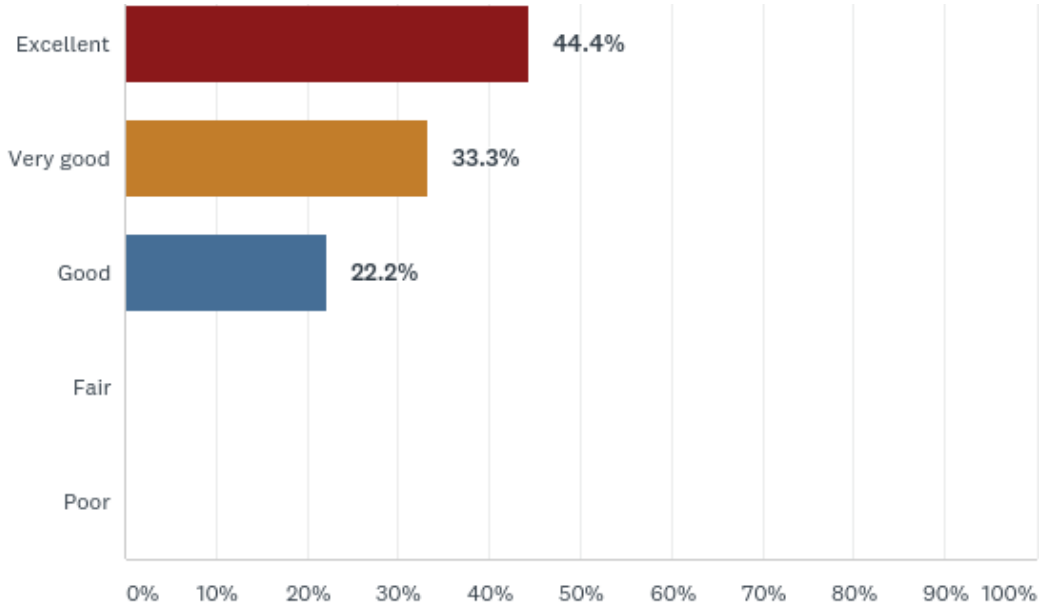
Q3. How would you judge the length of time spent introducing and discussing the description of the "Just Qualified" (or "Borderline Pass") candidate (approximately 45 minutes)?



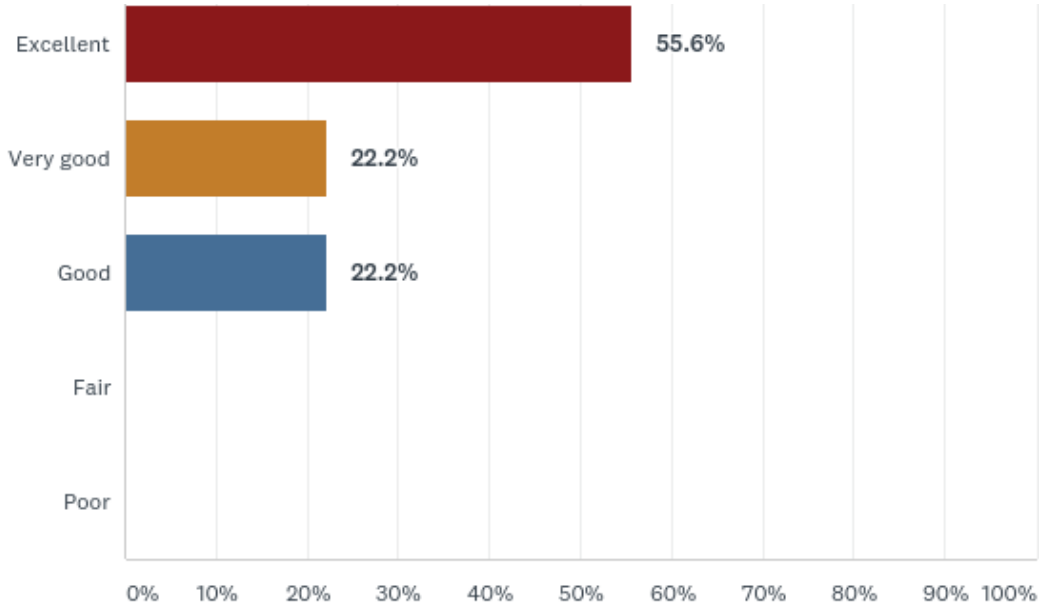
Q4. What is your impression of the length of training time you received for setting a pass score for the MCCQE Part II?



Q5. How clear was the information provided regarding the scoring procedures for the MCCQE Part II?



Q6. What is your overall evaluation of the training provided for setting a pass score for the MCCQE Part II?

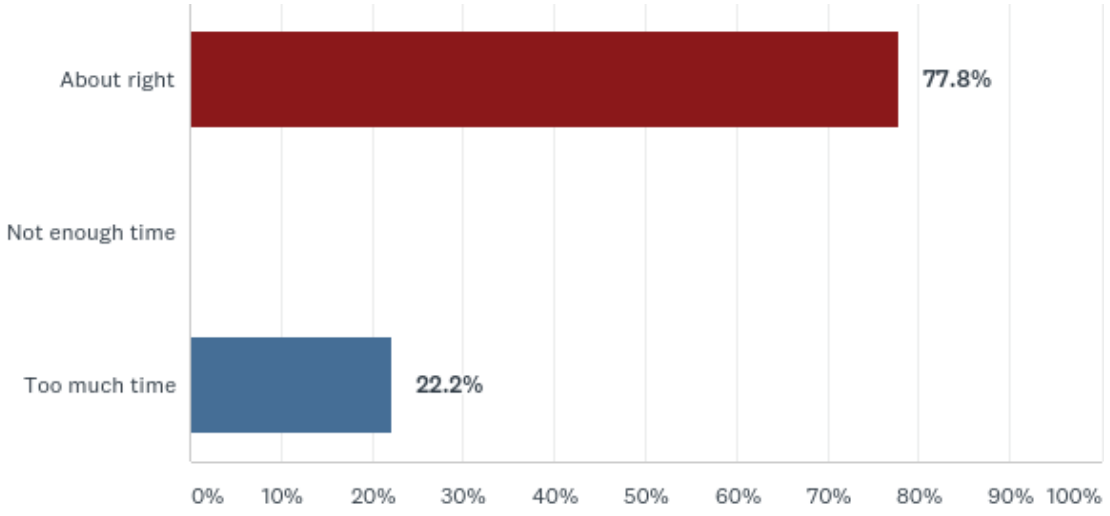


Q7. What factors influenced the ratings you made of the "Just Qualified" (or "Borderline Pass") candidate responses on the MCCQE Part II? Please select all that apply.

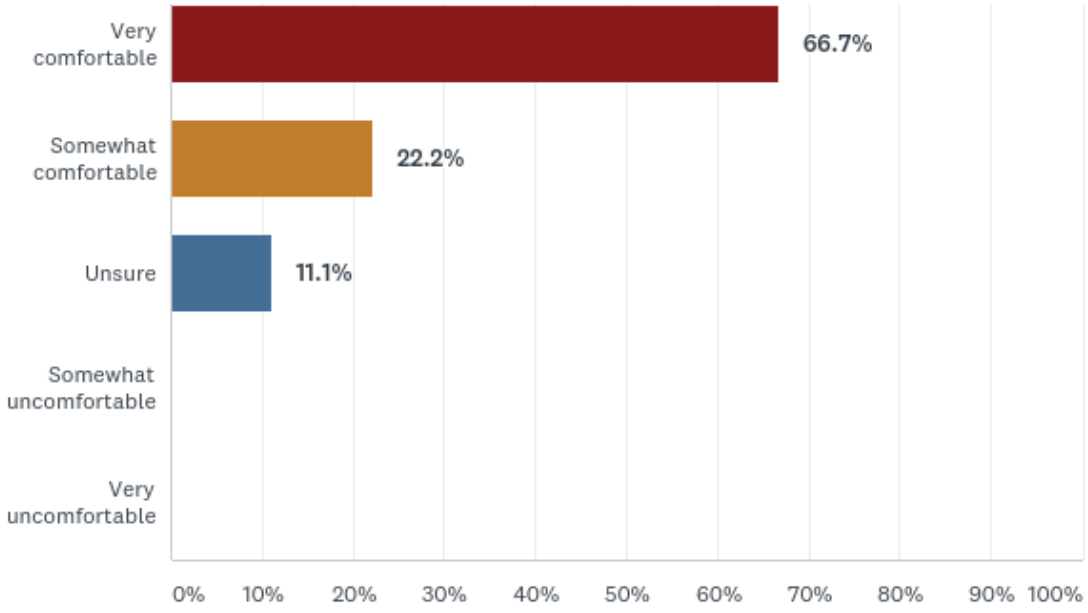
ANSWER CHOICES	RESPONSES	
The description of the "Just Qualified" or "Borderline Pass" candidate	88.9%	8
My perception of the difficulty of the stations or station components	66.7%	6
The scoring of the individual stations or station components	77.8%	7
The station statistics (e.g., candidate station scores)	55.6%	5
The statistical impact data provided before the final round	11.1%	1
Panelist discussions	88.9%	8
My experience in the field	77.8%	7
Knowledge and skills measured by the stations	77.8%	7
Other (please specify):	11.1%	1
Total Respondents: 9		

Other (please specify): *"The completion of critical items on the checklist. Also, the global impressions by the marker were important at the final outcome in these cases."*

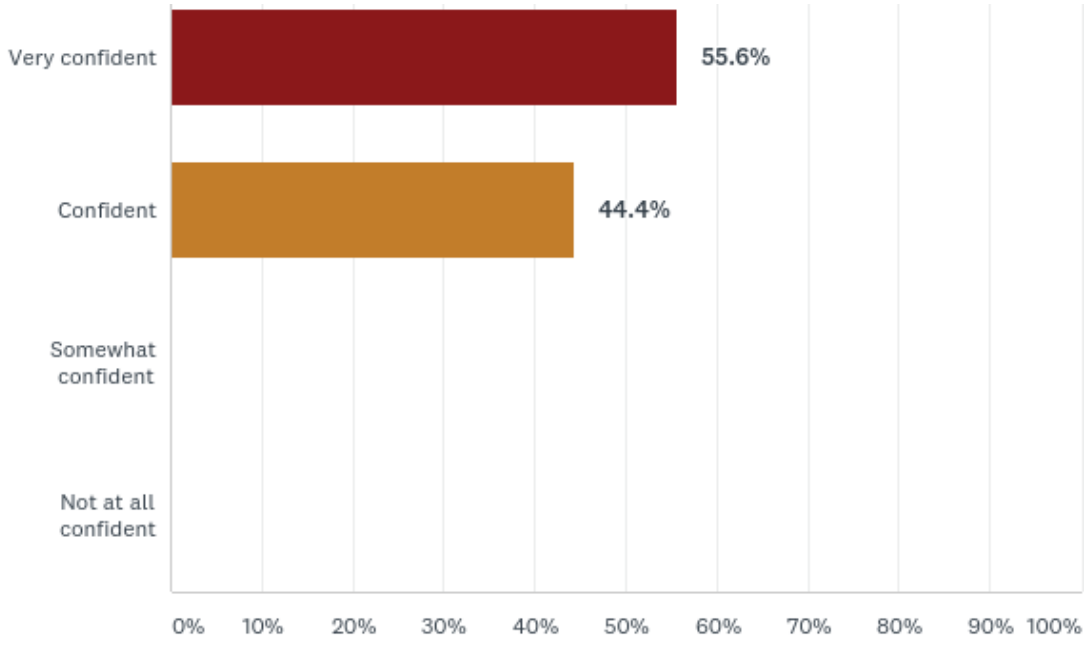
Q8. How would you judge the length of time provided for completing the ratings for each of the stations?



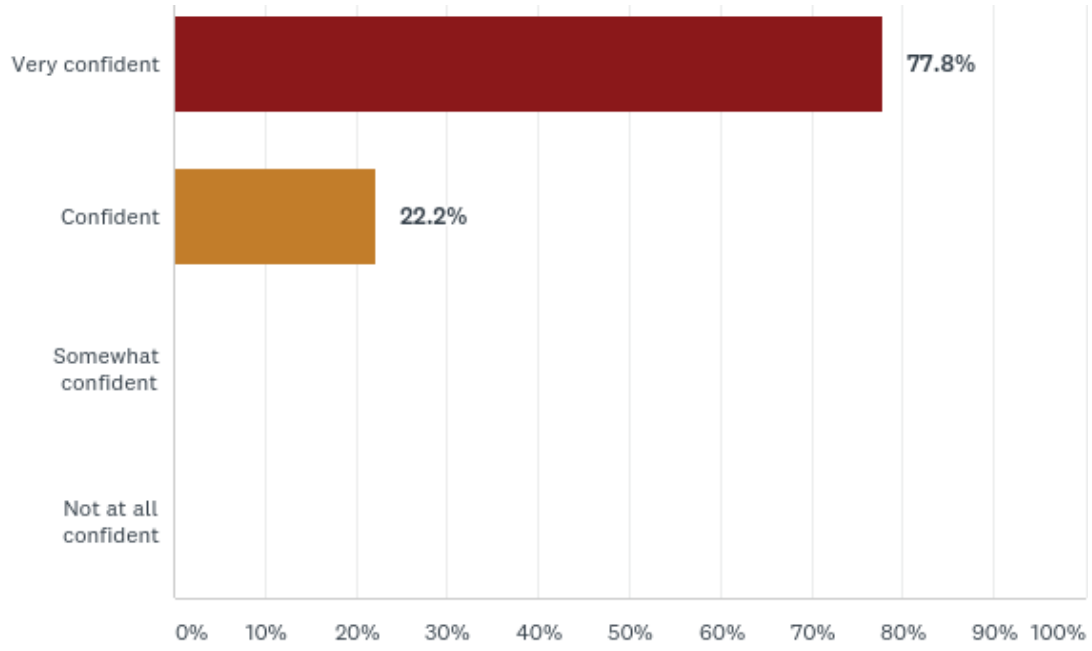
Q9. Overall, how did you feel about participating in group discussions conducted during the ratings process for each station?



Q10. What level of confidence do you have that the impact data and final discussion on the final afternoon helped the panel arrive at a defensible pass score?



Q11. What level of confidence do you have in the final recommended pass score for the MCCQE Part II?



Q12. How could the processes used for setting a pass score for the MCCQE Part II have been improved?

perhaps by reminding us to read the materials that were sent earlier
I believe this arrangement of two rounds of scoring is very appropriate.
no idea
maybe just show one borderline video per station
Perhaps editing videos to cut out down-time.
There is a question about whether borderline candidate videos could be reviewed as well regarding scoring and validation of the mark.
What was done in the second day to tell us not really to look at the scores in the scoring sheets but more on a general impression and feeling about what is more important than the other
possibly have more information given in advance
Shorten time of breaks and lunch to move the process through a bit faster.

Q13. Please provide any additional comments or suggestions about the setting of a pass score for the MCCQE Part II.

the preparation for this meeting was very thorough
interesting experience. I would do it again someday. hard to make it less mentally straining.
The confidence comes from the rigour of the process and the cumulative expertise of the panels. It was eye-opening to appreciate the differences between our observations of examinee performance on the videos and the ratings provided- very subjective. I think it is really important to be consider both items and the global ratings in deliberating an individuals assignment to borderline.
It was a very educational process and I am grateful to have been invited to get the opportunity to participate in this exercise.
Very well organised and welcoming members of the staff. Thank you !
n/a
Hofstee method was confusing as written -- could you present in a different way for a more visual learner?