National Assessment Collaboration

Technical report on the 2020 standard-setting exercise for the NAC Examination

Psychometrics and Assessment Services

Published: June 2022

MEDICAL COUNCIL OF CANADA
LE CONSEIL MÉDICAL DU CANADA

Table of contents

1. Introduction

Standard-setting is a critical component of any high-stakes assessment program, particularly for licensing/certification and selection decisions in the health professions. We need to assure the public that those passing a selection examination possess the required knowledge, skills, and abilities necessary for safe and effective patient care. Standard-setting is a process used to define an acceptable level of performance and to establish a cut score for one or more target levels of performance in the competency domains assessed by an examination. A rigorous and valid process for standard-setting should be adhered to for licensing examinations (Cizek, 2012). The processes and procedures implemented and the validity evidence gathered should be outlined to support the use of classification decisions (Kane 1994, Kane 1998). This report documents the processes, procedures, and results of a standard-setting exercise that was carried out virtually by the Medical Council of Canada (MCC) for the National Assessment Collaboration (NAC) Examination and administered in September 2020 during the COVID-19 pandemic.

The NAC Examination is a one-day exam that assesses the readiness of international medical graduates (IMGs) to enter a Canadian residency program. It focuses on assessing core abilities to apply medical knowledge, demonstrate clinical skills, develop investigational and therapeutic clinical plans, as well as demonstrate communication skills at a level expected of a medical graduate entering postgraduate training in Canada. The performance exam comprises a series of Objective Structured Clinical Examination (OSCE) stations, which depict various clinical scenarios including problems in medicine, pediatrics, psychiatry, surgery, obstetrics and gynecology, and preventive medicine and public health. The NAC Examination consists of 10 operational stations for the September 2020 exam session.

Candidates rotate through a number of OSCE stations, each of which involves a standardized participant (SP) and an examiner who observes the clinical encounter between the candidate and the SP and evaluates the candidate's performance using a standardized scoring instrument that includes a checklist of tasks, oral questions, and rating scales that are designed to assess up to seven clinical competencies (i.e., history taking, physical examination, data interpretation, investigations, diagnosis, management, and communication skills). The exam is scored in a compensatory way, which means all stations count equally towards the total score and a candidate's weak performance on some stations can be compensated by their strong performance on other stations. Each station score is calculated as a percent-correct score by dividing the sum of item scores by the maximum possible points for that station. Station scores are then averaged to obtain the total exam score. Score comparability across test forms is established through statistical linking.

The COVID-19 pandemic affected the way in which we deliver performance assessments such as the NAC Examination. To virtually deliver a high-quality and psychometrically defensible exam during COVID-19, we adjusted the structure of its content, format, and delivery, including, but not limited to

- modified physical examination stations (e.g., removing physical touching, having candidates tell the examiner what physical examination manoeuvres they would have performed and describe what findings they were trying to confirm),
- physical distancing and personal protective equipment (PPE) measures,

- removal exam day activities that require large groups to congregate,
- online delivery of candidate orientations and information, and
- replacement of pilot stations with wait stations.

Due to these changes to the NAC Examination administered under COVID-19 safety protocol, we did not statistically link the total scores to those on the previous exam sessions prior to the pandemic. Consequently, we did not apply the pass cut score that was previously established to the cohorts who took the NAC Examination during the pandemic. These changes warranted a new standard-setting exercise.

In addition, to support valid interpretation and use of exam results without advantaging or disadvantaging candidates over time, the MCC decided not to report total scores or subscores for the September 2020 administration. Total scores would not have been comparable with scores of previous NAC exam sessions, especially because March and September 2020 candidates were in the same Canadian Resident Matching Service (CaRMS) selection cycle. As such, it was recommended that two cut scores would be established and used for the September 2020 exam session: one for determining pass/fail status, the other for differentiating between pass and pass with superior performance based on the level of performance compared with a graduate from a Canadian medical school. Three categories of status were reported to candidates and programs: pass, fail, and pass with superior performance.

The NAC Examination is one of the requirements for IMGs applying for CaRMS. CaRMS is a very competitive process due to a large number of applicants competing for a small number of residency positions for IMGs each year. The reporting of the new pass with superior performance category could assist in the absence of reporting total scores for September 2020 and help support informed decision making for stakeholders who require a higher level of performance on the NAC Examination.

The MCC carried out the standard-setting exercise for the NAC Examination from October 19 to 21, 2020, with a panel of 20 physicians from across Canada. The exercise was conducted virtually via the Zoom videoconference application due to travel restrictions during the COVID-19 pandemic. The NAC Examination Committee (NEC) is responsible for overseeing the NAC Examination including the development and maintenance of the exam content and the approval of exam results. The purpose of the meeting was to arrive at two recommended cut scores for subsequent consideration and approval by the NEC.

In this report, we summarize the process, procedures, and results of the three-day virtual exercise that led to the recommendation of two cut scores for the NAC Examination.

2. Procedures

In this section, we describe how we selected the standard-setting methods, how we selected and assigned panelists to two subpanels, technology readiness, the materials we prepared and provided to the panelists for the three-day virtual meeting, and the events that took place during the three-day meeting.

2.1 Selecting a standard-setting method

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

- First, the NAC Examination is a criterion-referenced exam for which a cut score should be defined as an acceptable level of both knowledge and performance demonstrated given the intended use of the exam. Whether a candidate has achieved a certain performance level (e.g., pass, pass with superior performance) is determined by comparing an individual candidate's performance with a performance standard regardless of the performance of other candidates. Therefore, a criterion-referenced standard-setting method (e.g., Contrasting Groups or Borderline Group) is most appropriate for the NAC Examination.
- Second, the NAC Examination is a clinical performance exam consisting of a series of OSCE stations. Examinee-centred standard-setting methods (e.g., Contrasting Groups or Borderline Group) are most appropriate for performance assessments where expert judges review the performance of a group of examinees and provide global judgments as to the adequate level of performance (Cizek & Bunch, 2007). Examinee-centred methods are particularly well suited to the complex multidimensional nature of performance assessments. The Contrasting Groups method is an examinee-centred, criterion-referenced method that has been used for setting standards on high-stake licensure and certification OSCEs similar to the NAC Examination (e.g., United States Medical Licensing Examination [USMLE] Step 2 Clinical Skills).
- Third, the Contrasting Groups method is easier to use when setting two cut scores than the Borderline Group method, both from a time constraint and cognitive load perspective. Given the time constraints of a three-day meeting, we used the Contrasting Groups method to avoid repeating two rounds for each of the two cut scores (which would be required for the Borderline Group method). The two methods are similar in that they both require panelists to make holistic judgments on the overall performance of candidates by classifying them into two (or more) categories. In fact, the Borderline Group method can be viewed as a generalization of the Contrasting Groups method (De Champlain, 2013).
- Finally, we had experience setting two cut scores using the Contrasting Groups method in the 2019 in-person standard-setting exercise for the same exam. We adapted the procedure and process to accommodate the virtual meeting.

We also chose to complement the Contrasting Groups method with the Hofstee method. We describe the two methods below.

2.1.1 Contrasting groups method

The original Contrasting Groups method requires the use of total scores on the criterion of interest and to classify candidates into two categories (e.g., qualified vs. unqualified, masters vs. nonmasters). Then the total score on the exam in question that best discriminates between the two groups of candidates is selected as the cut score for that exam. Typically, the total score distributions of the two groups are graphed, and the cut score is set at the intersection (or midpoint of the intersection zone) of the two distributions. If false-positive and false-negative errors are a concern for your examination, it is recommended to move to the right or the left to minimize the error of greater concern (De Champlain, 2013; Downing et al., 2006).

In our application of the Contrasting Groups method (which is typical in the medical education field), we did not use the total score to classify candidates. Instead, we asked standard-setting panelists to review the score sheet of each candidate on each OSCE station of the NAC Examination, make a holistic judgment on the candidate's performance on that station, and rate it into one of three performance levels: level 1, *fail*; level 2, *pass*; and level 3, *pass with superior performance*. Each score sheet represented a performance profile on a station, and it included a candidate's scores on checklist items, oral questions, and competency rating scales recorded by an examiner during the exam session. The score distributions of the three groups of performance levels were plotted. The midpoint of the intersection zone between group levels 1 and 2 was selected as the cut score for pass and the midpoint of the intersection zone between group levels 2 and 3 was selected as the cut score for *pass with superior performance*.

A full description of how we used this method to set two cut scores is provided in the 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 Contrasting Groups method is "in touch with reality," we also used the Hofstee method to check for reasonableness from a political and cognitive 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 (or a range of possible cut scores) based on answers to the following four questions that panelists are asked to address based on their expertise and experience in the field, knowledge assessed, 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. In contrast, their 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 total score to define a cut score.

The Hofstee method is usually not used as a stand-alone method. For our purpose, we used it to define a range of cut scores to provide a "reality check" on the first cut score (i.e., for pass) set using the Contrasting Groups method. Our hope was that panelists' cut scores, using the Contrasting Groups method, would fall within the range of "acceptable" values as defined by panelists' answers to the four Hofstee questions (i.e., their "gut" estimates). 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 well-qualified panelists is an important step to ensure the validity of a standard-setting process and the resulting cut scores. In view of the inherent subjectivity of any standard-setting process, best practice dictates the selection of a panel that broadly represents the target subject matter expert population (i.e., physicians in Canada) with respect to background and educational characteristics (De Champlain, 2013).

In July 2020, the MCC sent out an email invitation to physicians across the country to solicit interest in participating in a virtual standard-setting exercise. This solicitation resulted in about 250 interested physicians, each of whom completed a demographic information form. In addition, physicians were also asked to provide information related to technical requirements (e.g., operating system, browser, microphone, webcam, internet speed, etc.). This was taken into consideration when selecting panelists to ensure that the selected physicians were equipped with appropriate technology for the virtual meeting. The original invitation letter and demographic survey are included in Appendix A.

Based on the demographic and technological information provided, the MCC selected 20 participants 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, (5) number of years in practice postresidency, (6) practice community, and (7) care setting. The main purpose of using two subpanels was to assess the generalizability of the cut scores across two parallel but independent groups of physicians (i.e., can we replicate the cut scores across two matched subpanels?); a critical source of validity evidence in support of the recommended cut scores. In addition, smaller subpanels may foster more discussions as they allow each participant more opportunities to share their perspective. Table 1 summarizes the demographic composition of the two subpanels.

Table 1: Demographic Information by subpanel for the 2020 standard-setting exercise

Demographic Information	Group	Subpanel 1	Subpanel 2	Total
Gender	Male	5	4	9
	Female	5	6	11
Geographic region	West	2	3	5
	Central	2	2	4
	Ontario	4	3	7
	Quebec	1	1	2
	Maritimes	1	1	2
Ethnic background	White	6	6	12
	Other	4	4	8
Specialty	Family Medicine	6	5	11
	Other specialties	4	5	9
Years in practice post-residency	1-10	6	5	11
	11-30	3	5	8
	≥ 31	1	0	1
Practice community	Urban	7	7	14
	Rural	3	3	6
Care setting	Hospital-based	3	5	8
	Community-based	7	5	12

2.3 Preparing materials for the standard-setting exercise

Preparing well is key to a smooth and successful standard-setting exercise. Preparation involved assembling materials for a training station to prepare panelists and allow them to practice using our IT standard-setting application and using the Contrasting Groups and Hofstee methods that were used for the operational stations. In addition, the crux of any standard-setting exercise was to define the target candidates for the proficiency level(s) targeted for the examination.

2.3.1 Technology readiness

Given the challenges of the virtual meeting for such a complex exercise, significant technology-related preparations were needed to ensure a successful virtual experience. Examples of our preparation work included:

- Ensuring panelists met the technical requirements including operating system (e.g., Windows 8 or higher), browser, internet speed, microphone, webcam, etc.
- Adapting an in-house IT standard-setting application to include presenting candidate score sheets to panelists along with collecting panelists' judgment data. To prepare, we conducted User Acceptance Testing of the tool with internal staff to ensure the application functioned as intended and no load issues identified.
- Setting up access to Zoom and securely accessing our IT standard-setting application through a Virtual Private Network

- Placing all documentation to be used for standard-setting securely in Box and instructing panelists to set up a Box account
- Conducting technical dry runs with panelists prior to the day of the meeting
- Ensuring that IT technical support staff were available throughout the three-day meeting

2.3.2 Materials for the training station

A pilot station from a previous exam session was used as the training station for training panelists on the standard-setting procedures. Three video-recorded candidate performances on the training station were selected, each demonstrating a performance that was a *fail, pass* or *pass with superior performance*. The materials provided for training and practice purposes included candidate instructions, scoring key, and score sheets for the three video performances, and 75 candidate score sheets on the training station. The score sheets were scanned into PDF files and loaded into the standard-setting IT application. They were presented to panelists for review and judgments in the order from the lowest to the highest station score.

2.3.3 Materials for the operational stations

One of the two test forms used for September 2020 exam session was used for standard-setting. A stratified random sample of 75 candidates were selected from the candidate pool who took this form based on their total exam scores to cover a wide range of total scores. For the 75 selected candidates, all their score sheets for each of the 10 operational stations were used to provide judgments for the standard-setting exercise. Each score sheet represented a candidate's station performance profile, and it included the candidate's results on checklist items, oral questions, and competency rating scales on a station as marked by an examiner during the exam session. The background materials for the 10 stations were prepared for panelists. For each station, this included candidate instructions, props (i.e., support materials such as a medical chart), scoring key, score sheets for two video performances (see below), and 75 candidate score sheets. The score sheets were scanned into PDF files and stored in the standard-setting IT application. For each station, the score sheets were presented in the order from the lowest to the highest station score to panelists for review and ratings. The order of candidate score sheets differed from station to station because candidate performances varied by station (e.g., the highest station score for a particular candidate on the first station was not necessarily the highest station score on subsequent stations).

Two video-recorded candidate performances on each station were prepared for the standard-setting exercise, one represented a *pass* and another represented *a pass with superior performance*. The videos were selected by physician subject matter experts from the candidate performances recorded in test centers.

2.3.4 Performance level definitions

A critical step in any standard-setting exercise is to define the target candidate for the proficiency level targeted by the examination. The NAC Examination is intended to assess clinical competence at the level of a graduate from a Canadian medical school who is about to enter residency training in Canada. For the purpose of setting two cut scores, it was necessary

to define three targets: one was a *fail* candidate, second was a *pass* candidate, the other a *pass with superior performance* candidate. These targets were defined in terms of competence required for entry into residency training in Canada.

The performance expectations for the candidate at the level of *fail, pass,* and *pass with superior performance* were previously defined by a different group of physicians through a one-day focus-group meeting in preparation for the 2019 standard-setting exercise. That group of physicians who were knowledgeable about medical education, resident selection, and who were familiar with the IMG population (see the *2019 Technical report on the standard-setting exercise for the NAC Examination*). The performance levels were defined in the context of three broad domains of physician activities: assessment/diagnosis, management, and communication. The definitions were adjusted to reflect the changes made to the September 2020 NAC exam administered under COVID-19 safety protocols and approved by the NEC. These definitions are included in Appendix B.

2.4 Premeeting training of panelists

To prepare panelists, we implemented the following activities:

1. We developed training modules on test security, an overview of the NAC Examination and an overview of standard-setting. The modules were delivered through a Learning Management System (LMS). The panelists were required to complete the modules prior to the day of the standard-setting exercise to gain a general understanding of the exam and standard-setting before the meeting. Some of the content was repeated during the meeting to reinforce concepts presented. In addition, all panelists were required to sign the Code of Business Conduct and Nondisclosure Agreement via LMS.

2. We conducted a brief 15-minute virtual technical dry run with each panelist to make sure they had no issue with logging into Zoom, accessing documents in Box, and accessing our IT standard-setting application.

3. We provided panelists access to the following documents in Box prior to the meeting: (a) an agenda for the meeting, (b) performance level definitions, and (c) two research papers that provided overviews of standard-setting (De Champlain, 2013; Downing et al., 2006). Panelists were encouraged to review these documents prior to the meeting.

2.5 Activities during the three-day virtual meeting

The agenda for the three-day meeting is provided in Appendix C. The morning of the first day was devoted to training the panelists. The remainder of the 2 ½ days included two rounds in which panelists provided ratings on the 10 operational stations. In between the initial round and final round, impact data and discussion occurred. We describe the training and two rounds of standard-setting next.

2.5.1 Training and practice

The success of any standard-setting exercise relies on extensive and robust training of standard-setting panelists. To this end, we devoted the morning of Day 1 exclusively to training the panelists. We began the meeting with a virtual introduction of facilitators and panelists, Zoom meeting etiquette, and an overview of the purpose of the meeting. We specifically told panelists that their task was to recommend two cut scores and that we would submit their recommendations to the NEC for consideration and approval.

To familiarize the panelists with the exam, we provided an overview of the NAC Examination including its purpose, intended test-taker population, blueprint, content specifications, station format, scoring, and score reporting. We also explained the adjustments made to the structure of the exam content and format for delivery under COVID-19 safety protocols. Next, we followed with an overview of the standard-setting, which included its purpose, process, selection and training of panelists, issues and challenges, criterion- and norm-referenced frameworks, and common methodologies for OSCEs.

2.5.1.1 Discussion on performance level definitions

As part of the training, we devoted 45 minutes to reviewing and discussing the performance level definitions. The panelists shared their thoughts on candidate performance: their characteristics, what they knew or were capable of doing, things they might have difficulty completing. Panelists were also asked to envision some *pass* and *pass with superior performance* candidates to identify what could distinguish *pass* from *fail* candidates, and *pass with superior performance* from *pass* candidates, etc. The MCC's Medical Education Advisor, who is also a practising physician, facilitated the discussion. The purpose of that activity was to calibrate the panelists to a common understanding and expectation of appropriate performance levels for the NAC Examination candidates. We instructed panelists to use these definitions to guide their judgment of candidate performance throughout the standard-setting exercise.

2.5.1.2 Practice using the training station

Using the training station, we provided step-by-step training on the standard-setting process. Specifically, we followed these steps:

Step 1: A Test Development Officer (TDO) introduced the station's objective and key features (critical elements) in resolving the clinical problem.

Step 2: A TDO reviewed the score sheet and scoring key.

Step 3: Panelists watched three video performances, one each for *fail, pass,* and *pass with superior performance.*

Step 4: A TDO facilitated a group discussion on station content and video performances.

Step 5: Panelists independently reviewed 75 candidate score sheets on the training station, rated their performance into three levels (i.e., level 1, *fail*; level 2, *pass*; level 3, *pass with superior performance*), and practiced entering their ratings in the standard-

setting IT application. As described in section 2.3.2, the score sheets were presented in the order from the lowest to the highest station score for that particular station.

The training, hands-on practice and thorough discussions were meant to help panelists develop a solid understanding of the performance-level definitions, standard-setting process, and their specific tasks.

2.5.2 Standard-setting exercise

Following the training, the panelists were asked to rate the performance of the 75 candidates for the 10 operational stations twice, during the initial and final rounds. Panelists were given access to all the materials in Box during the meeting. They were encouraged to use the "chat" function to pose questions and comments and facilitators were monitoring and addressing them throughout the meeting.

2.5.2.1 Initial round

For the initial round, we split the panelists into two subpanels and placed them in two breakout rooms for the rest of Day 1 and Day 2 of the meeting. A psychometrician and a TDO facilitated each subpanel. For each of the 10 operational stations, we followed the same five-step process described in section 2.5.1.2 for the training station except that in step 3, only two videos were presented (one for *pass*, the other for *pass with superior performance*).

The panelists were given as much time as needed to provide their ratings, initially the task took longer, but over time, the tasks were less time-consuming because the panelists were more familiar with the tasks, materials, process, and IT standard-setting application. The panelists provided ratings independently of other panelists, and there was no discussion of ratings during this part of the exercise.

After panelists had completed their ratings for all 10 stations by the end of Day 2, we asked them to provide and record answers to the four Hofstee questions as described in section 2.1.2 using the form provided in Appendix D. We applied the Hofstee method to the first cut score only (i.e., for *pass*). Specifically, we asked panelists to specify the highest and lowest cut scores as well as the highest and lowest failure rates that they thought would be reasonable for the NAC Examination based on their holistic judgment of the purpose, content, and intended test-taker population, as well as the intended use of the exam.

2.5.2.2 Presentation of impact data and discussion

We calculated the two cut scores (see section 2.5.2.4) by individual panelist, subpanel, and full panel using the ratings collected in the initial round. We also calculated the impact of the full panel's cut scores using candidate performance data on the NAC Examination from the September 2020 cohort. In addition, we used panelists' answers to Hofstee questions to define a range of "acceptable" cut scores for pass by individual panelist, subpanel, and full panel. Finally, we obtained the full panel results by averaging the results between the two subpanels.

Before the beginning of the final round on the morning of Day 3, we reconvened the two subpanels and presented the results from the initial round including

- Two cut scores by individual panelist (anonymized), subpanel, and full panel
- Impact of the two cut scores on the performance of first-time test takers: percentage of candidates who fell below the first cut score (i.e., for *pass*) and the percentage who fell below the second cut score (i.e., for *pass with superior performance*)
- Impact of the two cut scores on the performance of total test takers: percentage of candidates who fell below the first cut score and the percentage who fell below the second cut score
- Hofstee range of "acceptable" cut scores (for *pass*) by subpanel and full panel
- Afterwards, panelists discussed the results and impact data.

The results, impact data, and discussions helped to calibrate the panelists towards a better understanding of the process and potential consequences of their judgments. It also became clear to panelists why they needed to have a common understanding of the performance level definitions and to keep them in mind while providing ratings of candidate performance. Impact data, discussion, and their individual cut scores provided valuable feedback on the impacts of their ratings. For some panelists they would adjust a fair amount of their ratings on all stations, some panelists might make fewer changes if they were happy with their individual cut scores from the initial round.

2.5.2.3 Final round

We again split panelists into two subpanels in the final round and assigned them to two breakout rooms. The IT standard-setting application displayed their individual ratings from the initial round so panelists could make changes based on the feedback and discussion based on the impact data. Within each subpanel, the following two-step process was used for each station:

Step 1: A TDO provided a brief summary of the station.

Step 2: Panelists independently reviewed and provided their ratings (level 1, *fail*; level 2, *pass*; level 3, *pass with superior performance*) for each of the 75 candidate score sheets.

By this time, panelists were very familiar with the process, and they were told that only the final round results would be recommended to the NEC. Panelists' ratings from the initial round were presented on the same screen for their reference. When entering their ratings for the final round, they were able to change or keep the same ratings from the initial round. Again, we reminded them to keep in mind the purpose of the exam and performance level definitions when reviewing score sheets and making judgments.

After panelists completed their ratings for all 10 stations, we gave them a 40-minute break while staff members calculated the results and impact. We then presented the results and impact data from the final round to the full panel.

At the conclusion of the meeting, we asked panelists to provide feedback on the standard-setting exercise by answering an online survey anonymously.

2.5.2.4 Calculation of the two cut scores

In this section, we describe how we calculate the two cut scores for *pass* and *pass with superior performance*, respectively.

Cut score for *pass*:

We used each panelist's rating for each station to derive an individual panelists' station cut score, which was the midpoint between the maximum score of the group of candidates rated as fail and the minimum score of the group of candidates rated as pass as illustrated in Figure 1. We repeated this process for each of the 10 stations for each panelist. We then obtained each panelist's cut score for the total exam by taking the median of their 10 station cut scores and then took the median across all panelists' total exam cut scores in each subpanel to obtain a subpanel's cut score. Finally, we took the mean (same as the median) between the two subpanels' cut scores to obtain the full panel's total exam cut score for pass.

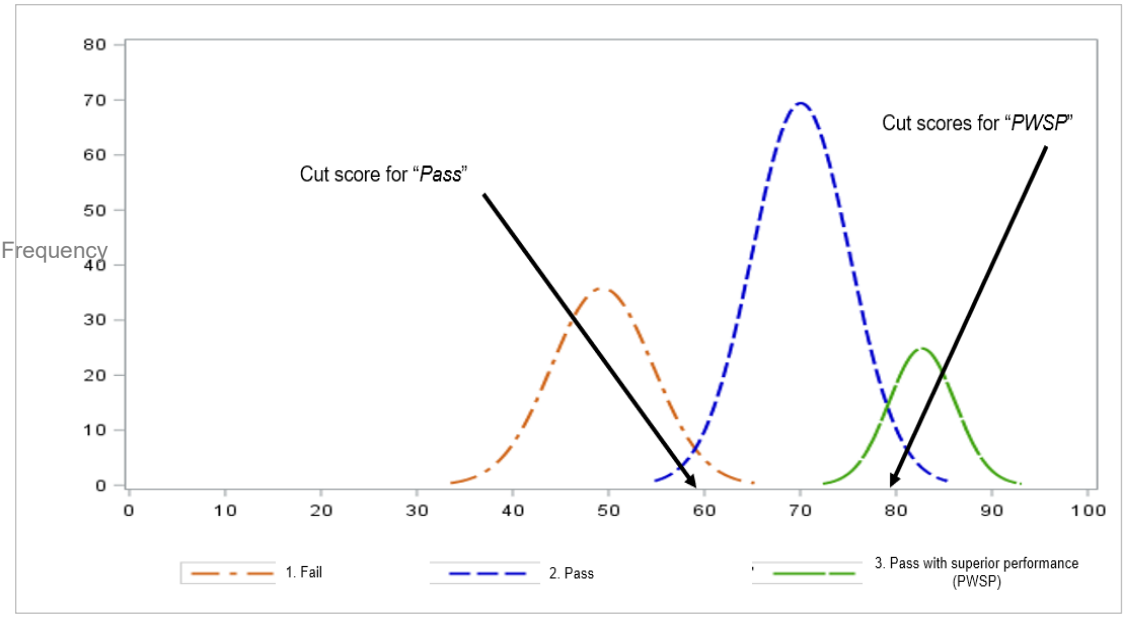


Figure 1: Illustration of cut-score calculations

Cut score for *pass with superior performance (PWSP)*:

The process for calculating the cut score for *PWSP* is the same as that for the *pass* except that we used the midpoint between the maximum score of the group of candidates rated as *pass* and the minimum score of the group of candidates rated as *PWSP*.

3. Results

In this section, we present results of the Contrasting Groups method, generalizability analyses, impact data, Hofstee results, and NEC's review and approval. Finally, we provide the results of the postsession survey.

3.1 Contrasting groups results

The cut scores were very similar between subpanels and between rounds; however, the variability across panelists decreased in the final round for both subpanels and the two subpanels converged in the final round. As indicated earlier, no total scores or subscores were reported for September 2020 exam administration. Only three categories of results were reported to candidates (i.e., *pass, fail, pass with superior performance*). The pass cut score was published on the mcc.ca website for the candidates that took the October 2021 onwards. The results of the initial and final rounds for the cut score between the fail and pass status are shown in Table 2.

Table 2: Summary of pass cut scores by round and subpanel for the 2020 standard-setting exercise

		Participants, No.	Min. cut score	Max. cut score	Median	SD
Initial round	Subpanel 1	10	27.8	56.9	**48.7**	9.4
	Subpanel 2	10	38.5	53.8	**43.8**	5.3
	Full panel	20			**46.3**	
Final round	Subpanel 1	10	36.1	61.5	**51.9**	7.8
	Subpanel 2	10	41.5	60.0	**48.5**	5.6
	Full panel	20			**50.2**	

3.2 Generalizability analysis 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 summary coefficients reflecting the level of dependability (D-coefficient) and generalizability (G-coefficient) that are analogous to the classical test theory's reliability coefficient. Multiple sources (commonly called facets) of error in a measurement can be estimated separately in a single G-analysis (e.g., persons or candidates, items, or in the case of OSCEs, stations, raters or panelists, and subpanel). The purpose of our analyses was to determine how much variance in the ratings was attributable to sources that are undesirable. We would want these sources to be as low as possible, such as panelists, subpanels, and stations, and how much variance was due to actual differences in candidate abilities (true score variance, which is desirable to separate candidates into different performance levels).

We conducted a G-analysis with three facets (*station, panelist,* and *subpanel*) in a *candidate* × *station* × (*panelist:subpanel*) design. In other words, the same 75 candidates were rated on the same 10 stations by panelists who were nested (assigned) to a specific subpanel (10 in each subpanel). We used the ratings panelists provided in the final round for these analyses. Table 3 shows the variance components for the panelists' ratings of candidate performance as well as each source of possible measurement error. These results from the G-analysis of the ratings obtained from the virtual standard-setting exercise are very similar to those we obtained in 2019 for the in-person standard-setting exercise for the NAC Examination.

The largest facet, not surprisingly, was the *candidate* × *station* interaction, which accounted for 58.7% of the total variance in panelists' ratings. This indicates that the rating of candidates (on the 1–3 scale) varied by station. This is commonly referred to as case specificity (Norman et al., 2006), typical of OSCEs, meaning that panelists' ratings of candidate performance on any station were specific to that station and do not necessarily generalize very well to other stations.

The second-largest facet was the *candidate* facet, which explained 13.3% of total rating variance, suggesting that candidates differed in their overall ability. This is akin to true score variance and indicates that panelists' ratings were able to separate out candidates, in terms of their performance levels. The third-largest effect was the station facet, which accounted for 5.9% of the total rating variance. This indicates panelists' ratings differed by station; therefore, the resulting cut scores would change slightly if a different set of stations were used in subsequent test forms (i.e., overall test form difficulty is dependent on the set of stations).

Because the panelists were nested within each subpanel, the *panelist* effect cannot be interpreted without the associated nested component of *subpanels*. The panelist-related effects were the next group of facet effects that were examined: *panelist:subpanel* accounted for 2.4% of total variance; *station* × *(panelist:subpanel)* explaining 1.2% of total variance and there was almost no variance attributed to *candidate* × *(panelist:subpanel)* facet. Together, approximately 3.6% of the total rating variance was due to the *panelist* nested within the *subpanel*. In other words, there was a small amount of variability in the resulting cut scores across panelists in both subpanels, mostly due to a few outliers. This justified our approach of using the median instead of the mean in each subpanel as their cut scores to minimize the effect of extreme values.

Next, we examined the effects related to *subpanel*. The *candidate* × *subpanel* and *station* × *subpanel* effects accounted for little rating variance (approximately 0.1%). These results indicate that there was a negligible amount of variance due to the two subpanels. As a matter of fact, the cut scores for the two subpanels were very close.

The G-coefficient and D-coefficient for the model "*candidate* × *station* × (*panelist:subpanel*)" were 0.69 and 0.67, respectively.

Table 3: Results of generalizability theory variance component estimates for the 2020 standard-setting exercise

Facet	*df*	SS	EMS	VCE	Total variance, %
Candidate	74	1797.30	24.29	0.08	13.3
Station	9	578.21	64.25	0.04	5.9
Subpanel	1	13.32	13.32	0.00	0.0
Candidate x station	666	5004.97	7.51	0.37	58.7
Candidate x subpanel	74	7.76	0.10	0.00	0.0
Station x subpanel	9	9.15	1.02	0.00	0.1
Candidate x station x subpanel	666	69.22	0.10	0.00	0.0
Panelist:subpanel	18	214.80	11.93	0.02	2.4
Candidate x (panelist:subpanel)	1332	151.35	0.11	0.00	0.0
Station x (panelist:subpanel)	162	113.90	0.70	0.01	1.2

Abbreviations: df, degrees of freedom; EMS, expected mean square; SS, sum of squares; VCE, variance component estimate.

In summary, the results of the G-analysis suggest that the ratings provided for this standard-setting exercise would generalize reasonably well if a different set of candidates, panelists, or subpanels were to be used, but less well if a different set of stations were to be used since most of the variance was associated with *candidate* × *station* and *station* facets. This means that the cut scores established for this exam were dependent on the set of stations used to set the standard and would necessitate that statistical linking be implemented to ensure score comparability across test forms (Kolen & Brennan, 2004). To address this, we conducted statistical linking to adjust for test form difficulty between the two test forms used for the September 2020 exam administered under COVID-19 safety protocols so that the same cut scores can be applied to both test forms. The same statistical linking process is applied to candidates taking the NAC Examination October 2021 onwards.

3.3 Impact data – pass rates

In Table 4, we present the pass rate for the Initial round and Final round for the First-time candidates and all candidates (or total) for the NAC Examination, September 2020. The overall pass rate is lower for the Final round compared with the Initial round because the pass cut score increased between the Initial round and the Final round.

Table 4: Pass rates from the September 2020 standard-setting exercise full panel cut scores, by round[a]

	Initial round, pass rate	Final round, pass rate
First-time candidates	92.6	87.0
Total candidates	92.7	88.1

[a] *Based on candidate performance data from the Standard-setting test form.*

3.4 Hofstee results

Table 5 summarizes the Hofstee results computed by averaging panelists' answers to the four Hofstee questions within each subpanel and for the full panel. As mentioned earlier, the Hofstee method was only used to define a range of "acceptable" cut scores for the pass cut score. The cut score range in the final round slightly decreased from that of the initial round for both subpanels and full panels.

The average Hofstee answers from the full panel in each round are plotted against a cumulative percentage of candidates who would fail at each point along the raw score scale using performance data of first-time test takers. Based on the Hofstee results in the final round, panelists felt that the cut score for *pass* should be no lower than 49.7% and no higher than 66.7%. Similarly, they indicated that the failure rate should be at least 18.0% but no higher than 38.5%.

Table 5: Summary of Hofstee method results[a] by round and subpanels from the 2020 standard-setting exercise

	Initial round				Final round			
	Acceptable cut score, %		Acceptable failure rate, %		Acceptable cut score, %		Acceptable failure rate, %	
	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.
Subpanel 1	47.8	73.8	18.5	42.7	49.7	65.0	16.6	37.9
Subpanel 2	49.7	67.4	16.2	31.1	49.4	68.4	19.3	39.1
Full panel	48.8	70.6	17.4	36.9	49.6	66.7	18.0	38.5

Abbreviations: Min., minimum; Max., maximum.

[a]The Hofstee method was used to define a range of "acceptable" cut scores for the pass cut score. The performance data used was from first-time test takers of the September 2020 exam session.

As indicated earlier, the Hofstee method was not our primary method for setting the standard for the NAC exam; it was used as a "reality check" of the standards set by using the Contrasting Groups method. The final cut score for pass fell within the range defined using the Hofstee method. This indicates that the final cut score for pass performance defined using the Contrasting Groups method was consistent with panelists' global judgment of what the cut score and failure rate should be from policy and cognitive perspectives.

3.5 Approval of cut scores

After a rigorous three-day standard-setting exercise, the panel of 20 physicians recommended the cut score for *pass* and the cut score for *pass with superior performance*. The recommended cut scores, impact data, and discussion information from this standard-setting exercise were presented and discussed by the NEC. As a result, the NEC approved new cut scores, which were, in turn, applied to the September 2020 NAC exam session.

3.6 Post-session survey

At the conclusion of the meeting, we asked panelists to provide feedback on the standard-setting exercise by answering an online survey anonymously. Eighteen panelists responded to the survey. Full results of the survey are presented in Appendix E. The following are the highlights of the survey results.

- Given the challenges of conducting the standard-setting exercise virtually, we invested significant effort in preparing panelists in advance of the meeting including detailed technical instructions, technical dry runs, and LMS training modules. The majority of panelists thought that the technical instructions were clear or very clear (94.4%), technical dry run was helpful or very helpful (83.3%), and all of them were comfortable or very comfortable with participating in the exercise virtually. In addition, they also found that premeeting training via LMS helpful or very helpful (88.9%).
- Central to the standard-setting exercise is the definition and description of the target candidate performance levels. Most panelists thought they benefited from a discussion on the *pass* and *pass with superior performance* levels and they found the discussion helpful or very helpful (77.8%). Most respondents thought they were clear or very clear (94.4%) about the performance level definitions as they began the standard-setting task in the final round.
- We devoted a significant amount of time and effort to training panelists on the standard-setting procedure to ensure a common understanding of what was expected of them before they engaged in the actual exercise. About 83.4% of panelists thought that the amount of training was adequate or very adequate. Most panelists thought that the hands-on practice was helpful or very helpful (83.3%). Overall, panelists thought that the training provided was excellent (22.2%), very good (44.4%), good (22.2%) or fair (11.1%).
- We solicited panelists' opinions on factors that influenced their judgment of candidate performance when reviewing score sheets. Multiple factors were considered from the most used to the least used: performance level definitions, candidate station score profile, panelist discussions, their experience with students/residents in the field, their perception of the difficulty of each station, knowledge and skills measured by each station, candidate's station scores, and the impact data presented to them after the initial round.
- At the end of the initial round, we presented impact data to show the consequences of their initial round cut scores. Panelists found the impact data and subsequent discussions to be helpful or very helpful (94.4%) in facilitating the panel to arrive at defensible cut scores.
- Finally, and most importantly, panelists indicated they were confident or very confident (88.8%) in the final recommended cut score for pass. They indicated confident or very confident (94.5%) in the final recommended cut score for *pass with superior performance*. None of the respondents indicated a lack of confidence.

4. Conclusions

Several findings highlight our confidence in the standard-setting process and the resulting cut scores.

- The two subpanels independently arrived at cut scores that were close in the initial round with absolutely no influence from each other. They converged closer in the final round though it is possible that, by this time, they might have been influenced by the initial round results, impact data, and discussions with other panelists. This provides evidence to support the careful selection and balanced assignment of the two subpanels as well as successful training to calibrate panelists to a common understanding of the performance level definitions and the standard-setting procedures. The similar cut scores by subpanel indicate that the cut scores can generalize across at least two matched subpanels.
- The G-analysis results provide additional validation of the results of this standard-setting exercise. Although there was some variability among individual panelists within each subpanel, the between-subpanel effect was virtually nil. This shows that, in general, the two subpanels performed in a similar manner, and more importantly, seemed to have a similar interpretation of the performance level definitions. These results are very similar to those in 2019 in-person standard-setting exercise for the NAC Examination.
- The cut score for pass defined by using the Contrasting Groups method was within the acceptable range defined by the Hofstee method based on panelists' holistic judgments. This indicates that the criterion-referenced cut score derived using the Contrasting Groups method is realistic and consistent with policy and practical considerations.
- The results of the postsession survey indicate a very positive experience from the panelists' point of view and the comprehensive training prepared them well to perform their tasks. Panelists expressed high confidence in the standard-setting process and the final recommended cut scores.

In summary, the similarity of the cut score by panel, G-analysis results, Hofstee results, and survey results all provide validity evidence that the standard-setting exercise was a thorough, rigorous, and valid process that meets best practice, and that the resulting recommended cut scores are defensible from both psychometric and policy perspectives.

The recommended cut scores were presented to the NEC on November 6, 2020, along with an overview of the standard-setting process, followed by the impact data. An additional option was presented to the NEC for the cut score between *pass* and *pass with superior performance*, that was one standard error lower as the number of candidates in the *pass with superior performance* category was fairly low. The NEC unanimously approved new cut scores for *pass* and *pass with superior performance* on the NAC Examination.

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). Routledge.

Cizek, G. J. and Bunch, M. B. (2007). *Standard-setting: A guide to establishing and evaluating performance standards on tests,* (pp.155-189). Sage Publications.

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). John Wiley & Sons.

Downing, S. M., Tekian, A. & Yudkowsky, R. (2006). *Procedures for establishing defensible absolute passing scores on the performance examinations in health professions education*. *Teaching and Learning in Medicine*, 18(1), 50-57.

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). Jossey-Bass.

Kane, M. (1994, September). *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.

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

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 Publications.

Appendix A: Invitation letter and demographic survey

MEDICAL COUNCIL OF CANADA | LE CONSEIL MÉDICAL DU CANADA

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

June 26, 2020

Dear potential panelist:

The purpose of this email is to invite you to express interest in serving as a panelist in a standard setting exercise for either the National Assessment Collaboration (NAC) examination or the Medical Council of Canada Qualifying Examination (MCCQE) Part II.

Both examinations are objective structured clinical examinations (OSCEs). The NAC examination assesses the readiness of international medical graduates for entry into Canadian residency programs. The MCCQE Part II assesses medical knowledge, skills and abilities expected of a physician in independent practice in Canada.

The COVID-19 pandemic has impacted how we deliver performance assessments such as OSCEs. We are making adjustments to how the NAC examination and the MCCQE Part II will be delivered so that we can deliver high-quality and psychometrically defensible exams and ensure a safe environment for staff and candidates. Given the changes to both examinations, the Medical Council of Canada (MCC) will be conducting standard-setting exercises to establish new cut scores for the NAC examination and the MCCQE Part II. To begin this process, the Psychometrics and Assessment Services directorate at the MCC is soliciting participation for two panels to recommend cut scores for the NAC examination and 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 these standard-setting exercises. 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.

Individuals who contributed to test development processes (e.g., content development, case writing, etc.) for the NAC examination and/or the MCCQE Part II in the past few years will not be selected as a panelist for the examination to which they had contributed; the validity of the pass score lies with a separation of test development processes from standard-setting processes.

Selected panelists will participate in the standard-setting exercise on **October 19-21, 2020 for the NAC examination** or **December 9-11, 2020 for the MCCQE Part II**. In normal circumstances, these exercises are conducted in-person. Due to COVID-19, we will conduct these meetings **virtually** via Zoom. Recognizing the challenges of working virtually, we have identified specific questions in the survey below regarding time-zones and technology availability/accessibility to provide the most seamless experience for all panelists. Panelists will be guided through a set of procedures to evaluate examination materials to set the cut score. Participants will receive an honorarium of $600 per day.

We hope that you will be interested in participating in one of our exercises. We ask that you complete the information survey by **July 13, 2020**, and tentatively reserve the standard-setting dates in your calendar. Your participation will be confirmed by July 24, 2020. 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, PAS

Medical Council of Canada (MCC) demographic survey for standard-setting exercise

The information requested below is being collected to help the MCC select two representative, pan-Canadian panels to recommend cut scores on the National Assessment Collaboration (NAC) Examination and the Medical Council of Canada Qualifying Examination (MCCQE) Part II. The standard-setting exercises will be held on:

- October 19-21, 2020 (NAC)
- December 9-11, 2020 (MCCQE Part II)

Due to COVID-19, these exercises will be held **virtually**. We want to make sure that we are taking the proper precautions to keep our panelists and staff safe and healthy.

Surveys must be submitted by **July 10, 2020**. Should you have any questions, please contact us at *research@mcc.ca*.

Demographics:

1. Please provide your full name and contact information (Name, email, and telephone 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 active unrestricted licence to practise with a Medical Regulatory Authority (MRA) in Canada?
    - No
    - Yes (please specify which province/territory):
5. Number of years in practice postresidency:
    - 0-2 years
    - 3-5 years
    - 6-10 years
    - 11-20 years
    - 21-30 years
    - More than 30 years

6. Have you had experience supervising students/residents?
    - No
    - Yes

7. How recently have you supervised students/residents?
    - 0-2 years
    - 3-5 years
    - 6-10 years
    - 11-20 years
    - 21-30 years
    - More than 30 years

8. Are you actively supervising students/residents?
    - No
    - Yes (please specify how often and how many students/residents you typically supervise each year): _____

9. 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

10. Number of years supervising International medical graduates (IMGs):
    - 1-5 years
    - 6-10 years
    - 11-20 years
    - 21-30 years
    - More than 30 years
    - I have no experience supervising IMGs

11. Have you ever participated in an MCC test committee or content development workshop?
    - No
    - Yes (please specify the activity and when): _____

NOTE: Being a test committee member or content development workshop participant is not a requirement to participate in the standard-setting exercise.

12. Have you ever been an examiner for the NAC Examination or MCCQE Part II? Please select all that apply.
    - I have been a NAC exam examiner
    - I have been an MCCQE Part II examiner
    - I have done both
    - I have not done either

    NOTE: Being an MCC examiner is not a requirement to participate in the standard-setting exercise.

13. Have you participated in a candidate preparatory course from a third party (i.e., not offered by the MCC) in preparation for the NAC exam or the MCCQE Part II within the last three years?
    - No
    - Yes (please specify the activity and when): _____

14. Where did you complete your postgraduate medical training?
    - Canada
    - Other (please specify): _____

15. Region of the country in which you currently practice:
    - Alberta
    - British Columbia
    - Manitoba
    - New Brunswick
    - Newfoundland and Labrador
    - Northwest Territories
    - Nova Scotia
    - Nunavut
    - Ontario
    - Prince Edward Island
    - Quebec
    - Saskatchewan
    - Yukon

16. First language:
    - English
    - French
    - Other (please specify): _____

17. Primary language of your medical practice:
    - English
    - French
    - Other (please specify): _____

18. Gender:
    - Female
    - Male
    - Prefer to self-describe: _____

19. Ethnicity:
    - Caucasian
    - Indigenous
    - Other group (please specify): _____

20. Medical specialty:
    - Pediatrics
    - Internal Medicine
    - Psychiatry
    - Obstetrics and Gynecology (OBGYN)
    - Surgery
    - Family Medicine
    - Other (please specify): _____

21. Type of community in which you primarily work:
    - Urban
    - Rural

22. Type of care setting in which you primarily work:
    - Hospital-based setting
    - Community-based setting

23. I am interested in and fully available to participate in the following standard-setting exercises (please select all that apply):
    - NAC exam (October 19-21, 2020 – three days)
    - MCCQE Part II (December 9-11, 2020 – three days)
    - I am interested and available for both exams

24. Do you have a preference for one standard-setting exercise over the other?
    - NAC exam (October 19-21, 2020 – three days)
    - MCCQE Part II (December 9-11, 2020 – three days)
    - I have no preference

Technology:

1. What time zone do you reside?
    - Newfoundland time zone
    - Atlantic time zone
    - Eastern time zone
    - Central time zone
    - Mountain time zone
    - Pacific time zone

2. What type of laptop/desktop do you have access to for this virtual meeting (cell phones and tablets cannot be used given the software/security requirements for this meeting)?
    - Windows 10
    - Windows 7
    - Apple
    - Other (please specify) (text box) _____

3. What browsers do you have access to for the virtual meeting?
    - Internet explorer 11
    - Firefox 27
    - Chrome 30
    - Safari 7
    - Other (please specify) (text box) _____

4. Do you have access to the virtual meeting? (check all)
    - Microphone
    - Webcam

5. What internet speed do you have access to for the virtual meeting?
    - Less than 3.0 Mbps up and down
    - Between 3.0 Mbps and 5.0 Mbps up and down
    - Greater than 5.0 Mbps up and down

6. Can you install software required for a Virtual Private Network (VPN) connection on the laptop/desktop for the virtual meeting?
    - Yes
    - No
    - Other (please specify): _____

Appendix B: Performance level definitions

NAC Examination standard-setting definitions

National Assessment Collaboration

PERFORMANCE LEVELS / COMPETENCIES	FAIL	PASS	PASS WITH SUPERIOR PERFORMANCE
PERFORMANCE LEVELS	**The candidate is *not* qualified to enter residency training.** The deficiencies are such that the candidate may put the patient at risk, or the candidate may not ensure the patient's basic needs are met.	**The candidate is qualified to enter residency training.** The deficiencies are such that the candidate does not put the patient at risk, and the candidate ensures the patient's basic needs are still met.	**The candidate is qualified to enter residency training and has demonstrated a superior performance.** The candidate mostly provides patient-centred, safe care.
COMPETENCIES – Assessment and Diagnosis	The candidate is often unable to gather the patient's essential information (through history taking, physical examination and laboratory data). The candidate's information gathering is disorganized, and the information they collect often lacks coherence, is missing critical details, or it contains critical details but has gaps in linking those details together. The candidate often lacks the knowledge to respond appropriately to information, and the candidate is often unable to synthesize information to formulate an appropriate differential diagnosis.	The candidate is able to gather most of the patient's essential information (through history taking, physical examination and laboratory data), but some aspects of their information gathering may be disorganized. The candidate may lack the skill to consistently develop a clear definition of the patient's problem. The candidate's misinterpretation of information or gaps in their knowledge or information gathering may affect the breadth and depth of their differential diagnosis.	The candidate is able to gather most of the patient's essential information in an organized and focused manner (through history taking, physical examination and laboratory data). The candidate mostly demonstrates the skills needed to develop a clear definition of the patient's issue and to develop a prioritized differential diagnosis.
COMPETENCIES – Management	The candidate has inconsistent and unpredictable management strategies for common, acute and emergent illnesses, and the candidate often lacks knowledge of treatment options. Their management plan is not patient-centred.	The candidate has basic management strategies for common, acute and emergent illnesses, but the candidate may lack more specific knowledge of treatment options. Their management plan has elements that are patient-centred.	The candidate mostly has appropriate strategies for managing common, acute and emergent illnesses. Their management plan is mostly patient-centred.
COMPETENCIES – Communication	The candidate may not communicate clearly with the patient or with the health care team. The candidate often does not respond to the patient's verbal and non-verbal cues or is often not empathetic or caring. In communicating with others, the candidate may be disrespectful and may demonstrate biases (e.g., gender, religious, sexual orientation or racial).	The candidate is generally able to communicate clearly with the patient and to summarize findings and plans with the health care team. The candidate often responds to the patient's verbal and non-verbal cues and is often empathetic and caring, although they are not always consistent. In communicating with others, the candidate is respectful and does not demonstrate biases (e.g., gender, religious, sexual orientation or racial).	The candidate communicates clearly with the patient, articulates clinical reasoning and summarizes findings and plans with the health care team. The candidate consistently responds to the patient's verbal and non-verbal cues and is empathetic and caring. In communicating with others, the candidate is respectful and is genuinely accepting of others.

Appendix C: Standard-setting meeting agenda

Virtual NAC Standard-setting exercise
October 19-21, 2020

Day 1: Monday, October 19

Time	Item
09:45 a.m.	Log into Zoom
10:00 a.m.	Welcome and introductions
10:25 a.m.	Review of agenda and objectives
10:30 a.m.	Overview of the NAC Examination
10:50 a.m.	Overview of standard setting
11:15 a.m.	Discussion on performance level definitions
12:00 p.m.	**Break (10 minutes)**
12:10 p.m.	Training and practice
13:35 p.m.	**Lunch (40 minutes)**
14:15 p.m.	Station T01 (Initial round)
15:55 p.m.	**Break (10 minutes)**
16:05 p.m.	Station T02 (Initial round)
17:30 p.m.	Station T03 (Initial round)
18:45 p.m.	**Wrap-up of day 1**

Day 2: Tuesday, October 20

Time	Item
09:45 a.m.	Log into Zoom
10:00 a.m.	Recap of day 1
10:05 a.m.	Station T05/T06 (Initial round continued)
12:20 p.m.	**Break (10 minutes)**
12:30 p.m.	Station T07 (Initial round continued)
13:35 p.m.	**Lunch (40 minutes)**
14:15 p.m.	Station T09/T10 (Initial round continued)
16:25 p.m.	**Break (10 minutes)**
16:35 p.m.	Station T11/T12 (Initial round continued)
18:35 p.m.	Hofstee method
18:50 p.m.	**Wrap-up of day 2**

Day 3: Wednesday, October 21

Time	Item
09:45 a.m.	Log into Zoom
10:00 a.m.	Present initial round results
10:15 a.m.	Sub-panel discussions (split into breakout rooms)
10:30 a.m.	Full panel discussions
10:45 a.m.	**Round 2 exercise** Station T01/T02 (Final round)
11:55 a.m.	**Break (10 minutes)**
12:05 p.m.	Stations T03/T05/T06 (Final round)
13:50 p.m.	**Lunch (40 minutes)**
14:30 p.m.	Stations T07/T09/T10/T11/T12 (Final round)
17:10 p.m.	Hofstee method
17:20 p.m.	**Break (40 minutes)**
18:00 p.m.	Present final results
18:15 p.m.	Post-session survey
18:25 p.m.	**Final wrap-up of day 3**

Appendix D: Hofstee form

Panelist: _____ Subpanel: _____

Round: Initial

Given the purpose of the exam, please specify a range of acceptable pass scores **based on content consideration** (between 0% and 100%)

1. What is the **highest** percentage pass score that would be acceptable? _____
2. What is the **lowest** percentage pass score that would be acceptable? _____

Given the purpose of the exam, please specify a range of acceptable failure rate **based on political consideration** (between 0% and 100%)

3. What is the **maximum** acceptable failure rate? _____
4. What is the **minimum** acceptable failure rate? _____

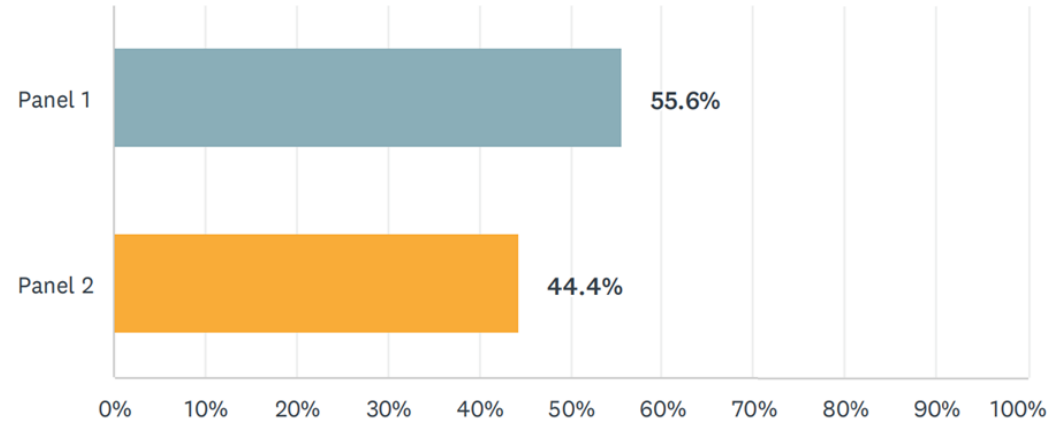
Round: Final

1. What is the **highest** percentage pass score that would be acceptable? _____
2. What is the **lowest** percentage pass score that would be acceptable? _____
3. What is the **maximum** acceptable failure rate? _____
4. What is the **minimum** acceptable failure rate? _____

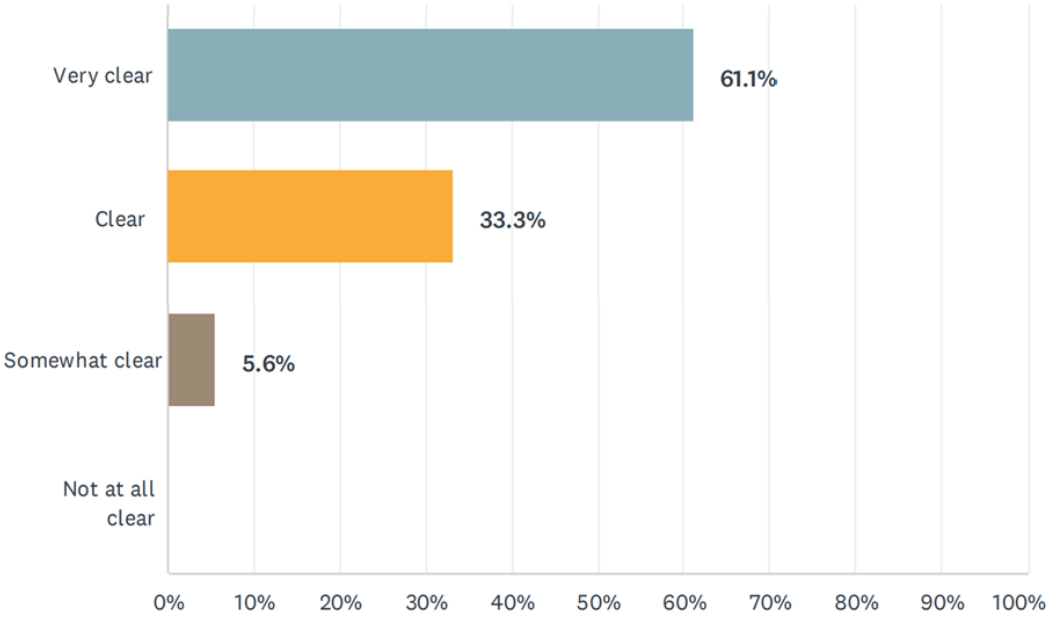
Your collective answers to the four questions will be used to define a range of acceptable pass scores that will be used to check the reasonableness of the cut score defined using the Contrasting Groups method.

Appendix E:
Summary of responses to post-meeting survey

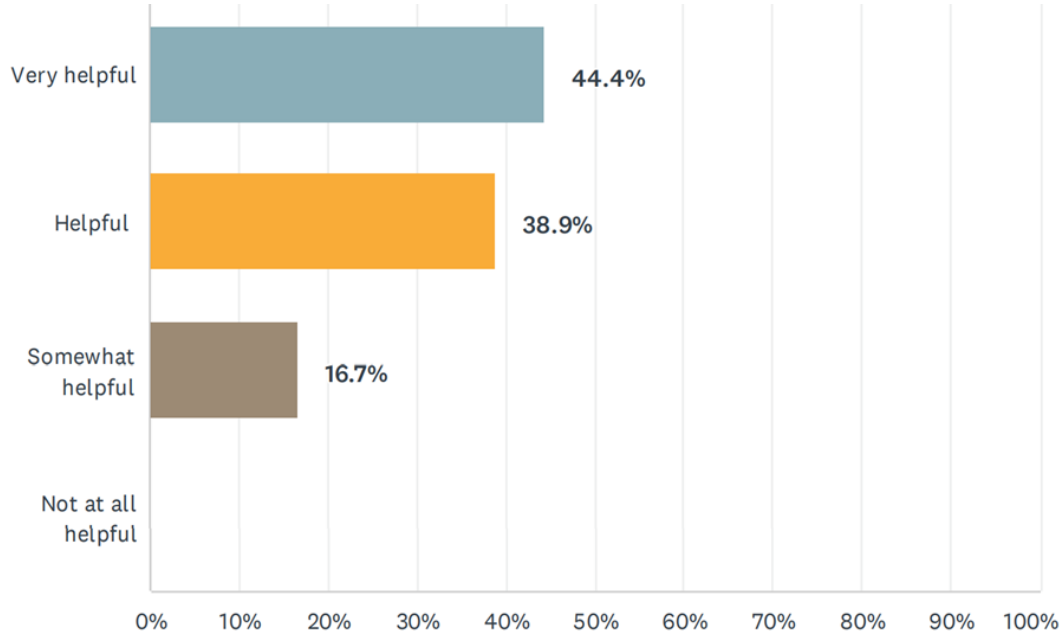
1. Which panel were you assigned for the standard-setting exercise?



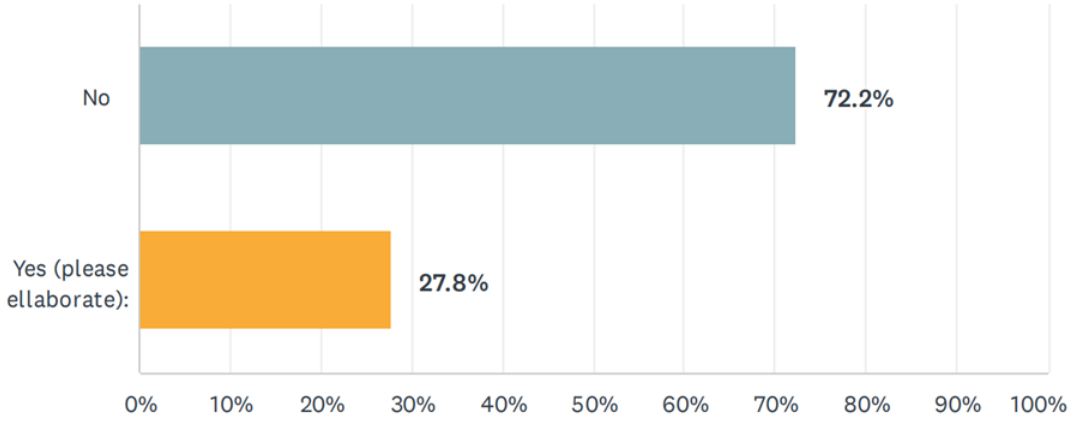
2. Were the technical instructions clear to you (BOX, Zoom, standard-setting rating tool)?



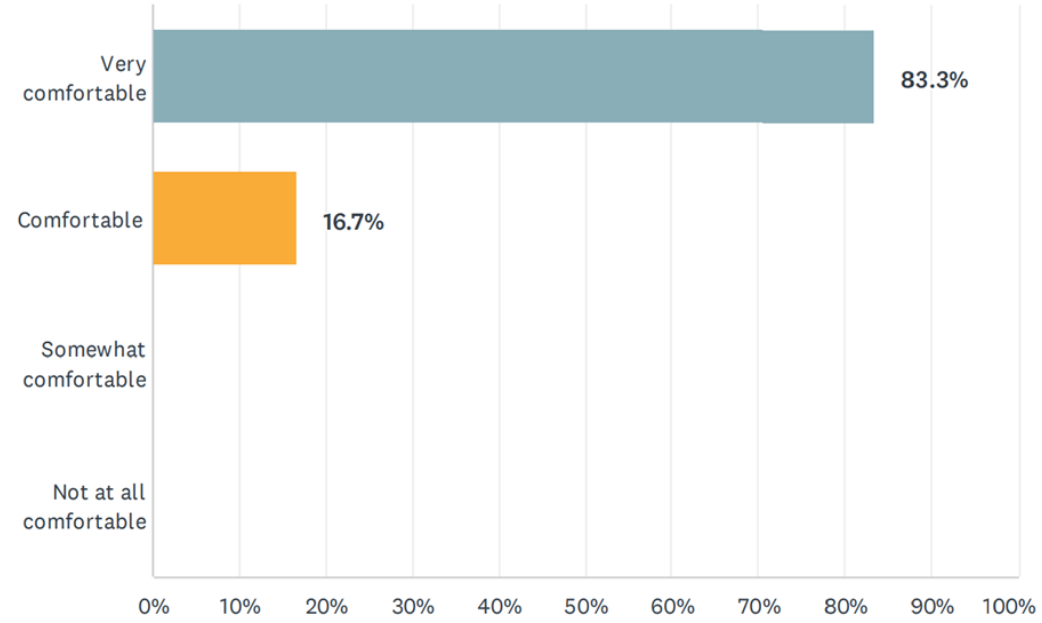
3. Was the premeeting technical dry-run helpful to you?



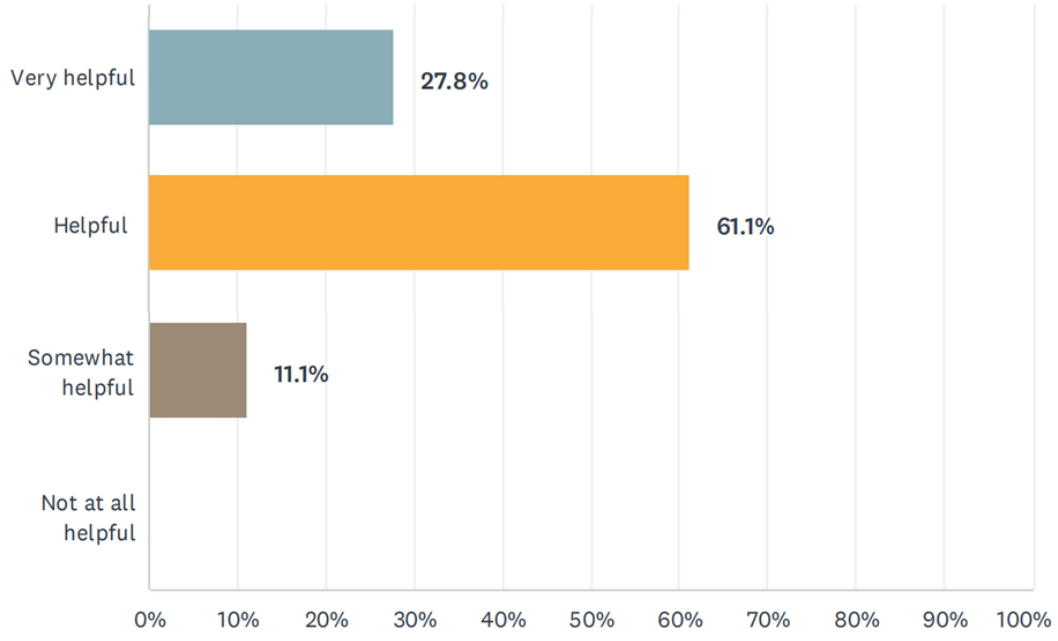
4. Did you encounter any technical difficulties during the meeting?



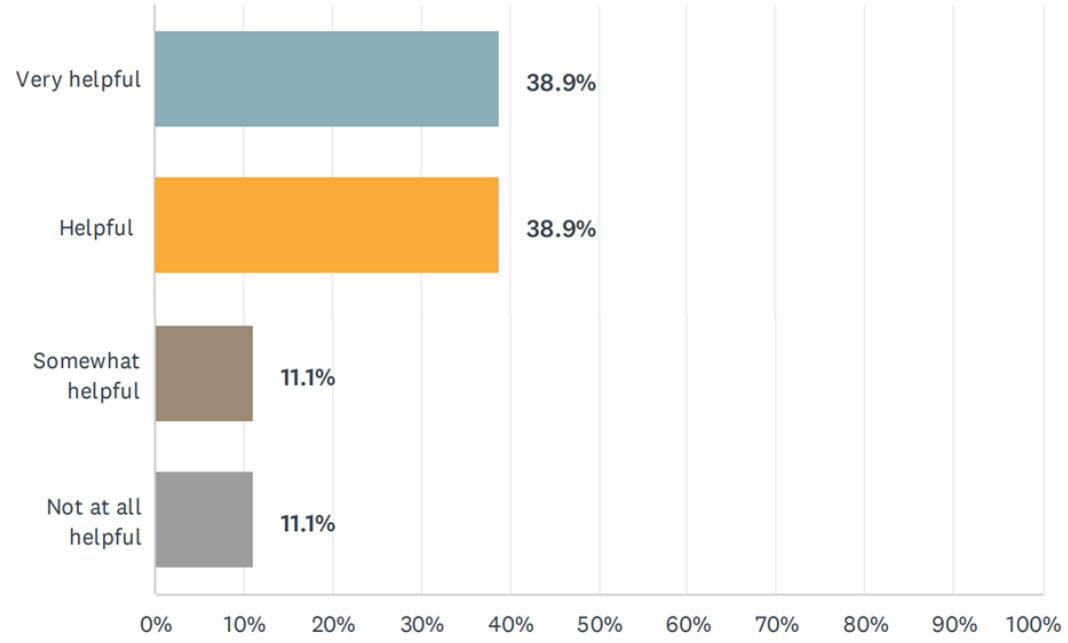
5. How comfortable were you with participating in the standard-setting exercise virtually?



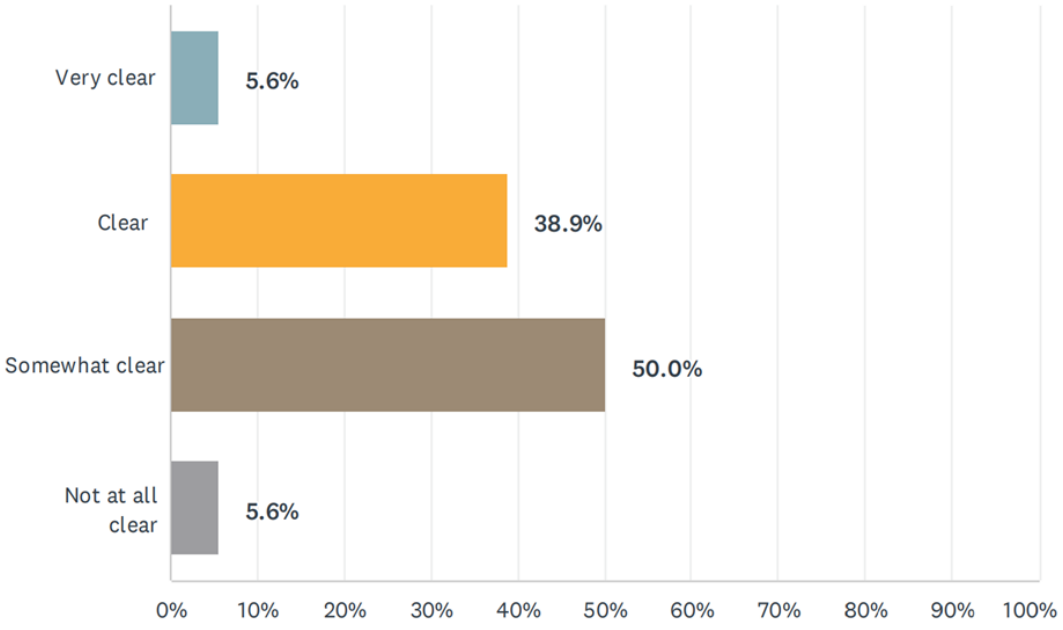
6. Did the premeeting modules in Learning Management System (LMS) prepare you well for the virtual meet?



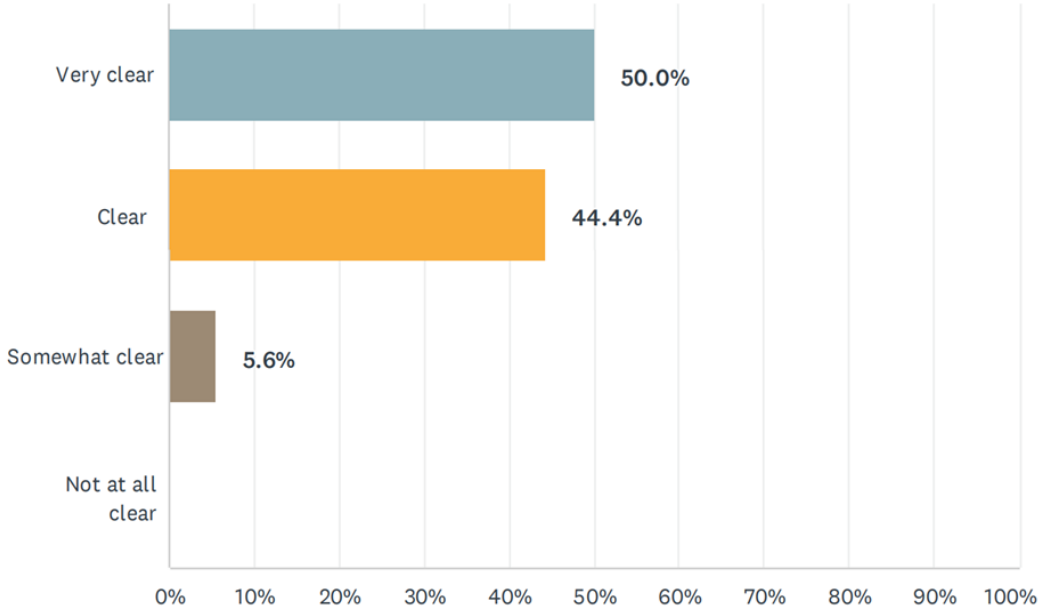
7. During the training on Day 1, how helpful was the discussion on the definitions for “pass candidate” and “pass with superior performance candidate” for the NAC Examination?



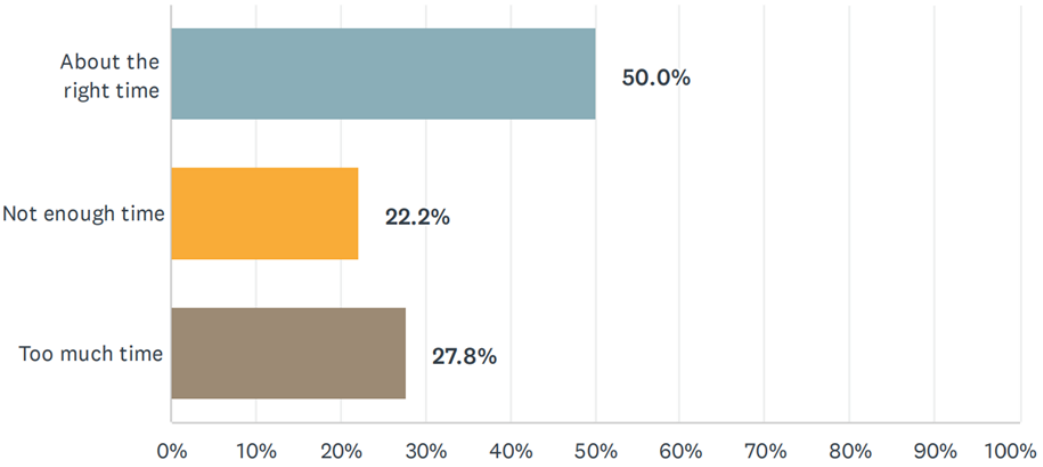
8. Following the training on Day 1, how clear was your understanding of the descriptions of the “pass candidate” and the “pass with superior performance candidate” for the NAC Examination as you began the task of setting cut scores in the initial round??



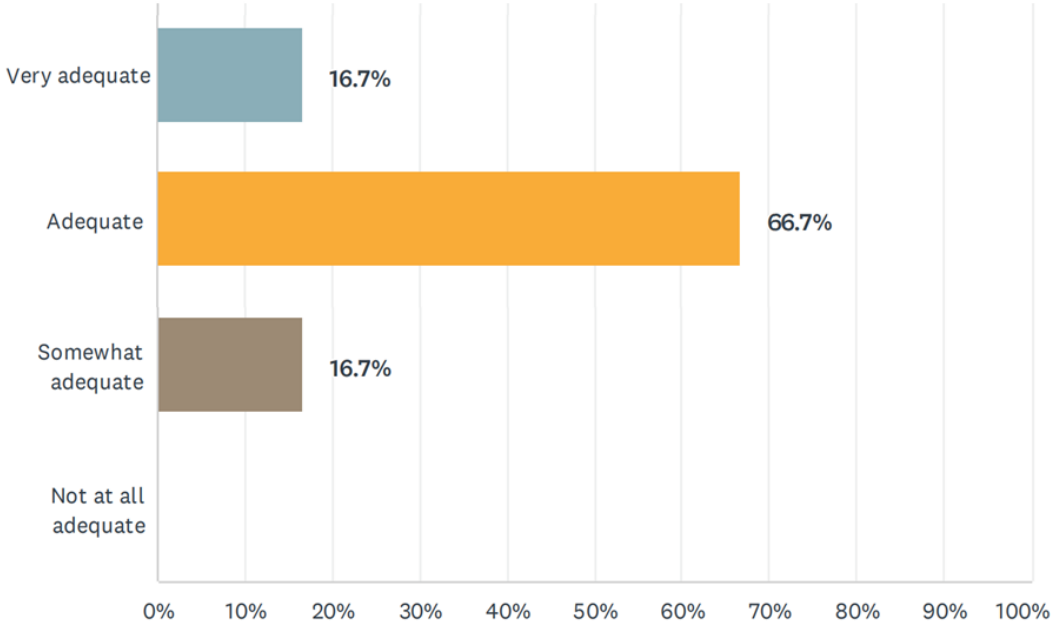
9. On Day 3, how clear was your understanding of the descriptions of the “pass candidate” and the “pass with superior performance candidate” for the NAC Examination as you began the task of setting cut scores in the final round?



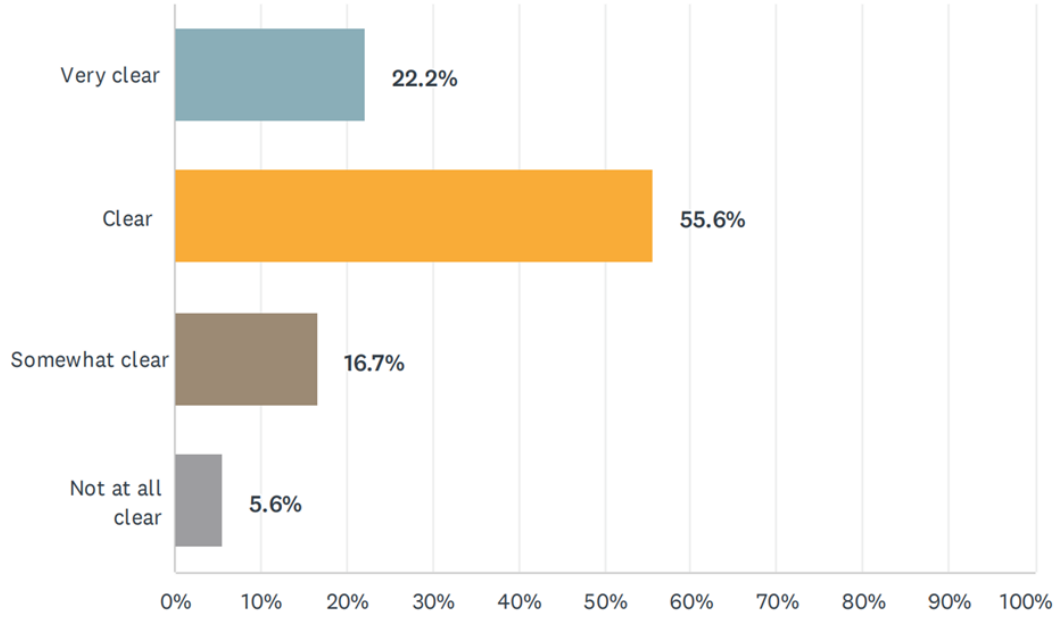
10. How would you judge the length of time spent introducing and discussing the definitions of the “pass candidate” and the “pass with superior performance candidate” (approximately 45 minutes)?



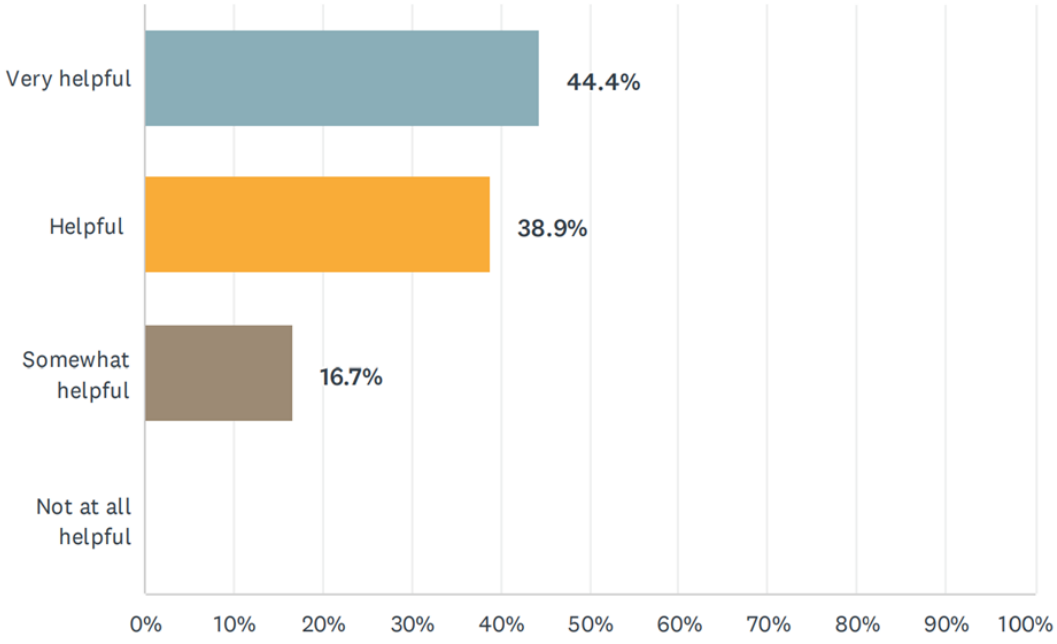
11. During the virtual meeting, what was your impression of the amount of training you received on setting the cut scores?



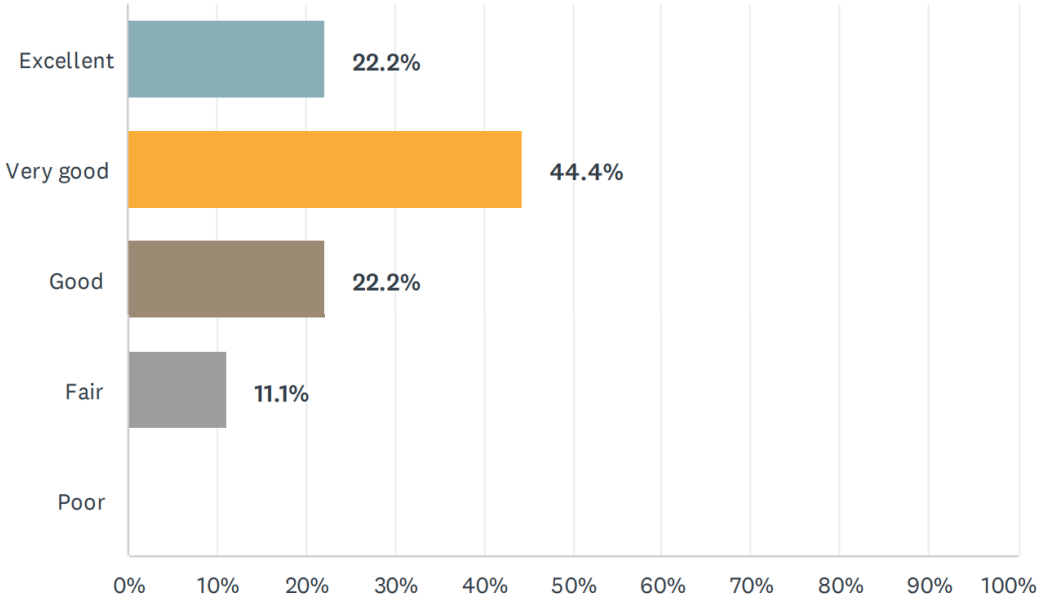
12. How clear was the information provided regarding the scoring procedures for the NAC Examination?



13. How helpful was the practice session for using the standard-setting application?



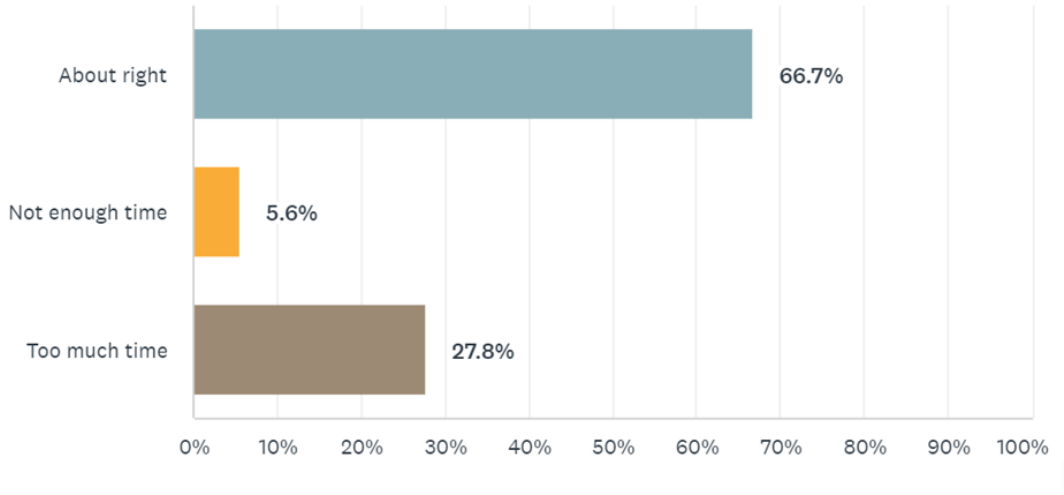
14. What is your overall evaluation of the training provided for setting cut scores for the NAC Examination?



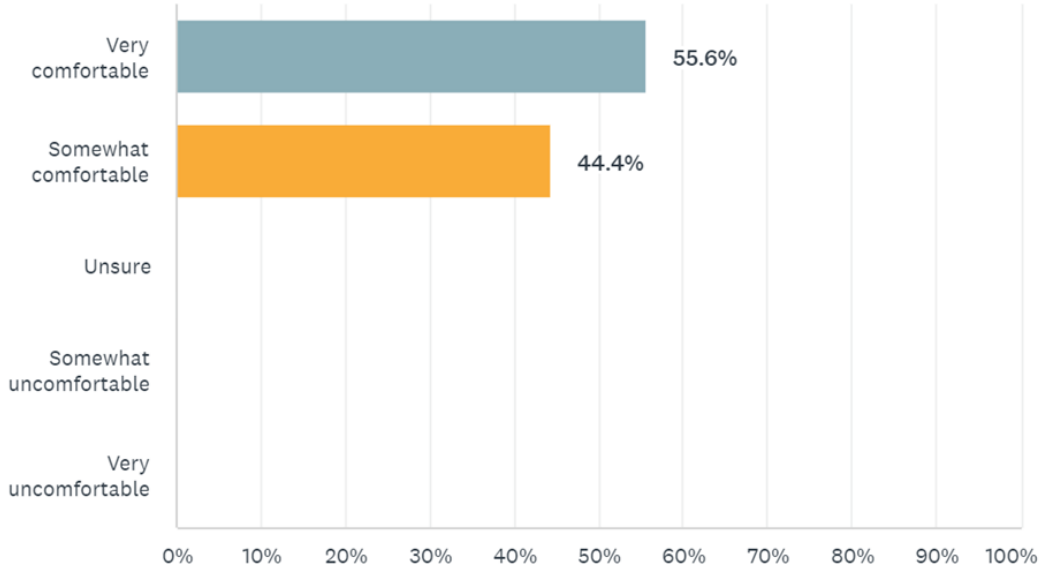
15. What factors influenced the ratings (i.e., 1, 2, 3) you made based on candidate score sheets on the NAC exam? (Select all that apply)

ANSWER CHOICES	RESPONSES	
Descriptions of the “pass candidate” and the “pass with superior performance candidate”	77.78%	14
My perception of the difficulty of each station	83.33%	15
Candidate's score profiles (on checklist items, oral questions, and rating scale items)	94.44%	17
Candidate station scores	77.78%	14
The impact data provided before the final round	77.78%	14
Panelist discussions	72.22%	13
My experience in the field	66.67%	12
My experience with students/residents in the field	66.67%	12
Knowledge and skills measured by each station	44.44%	8
Total Respondents: 18		

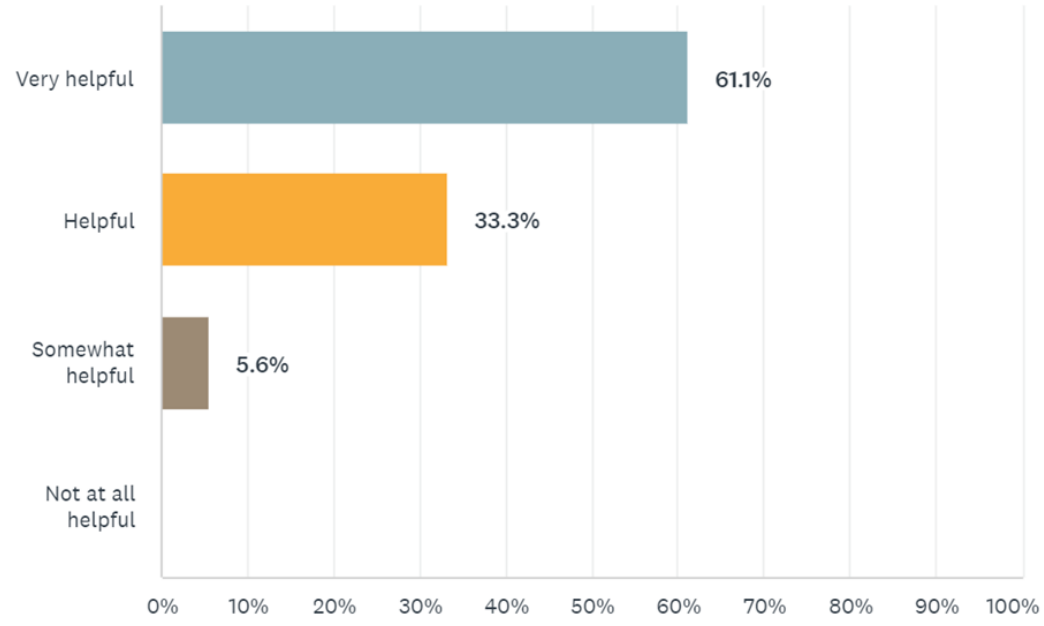
16. How would you judge the length of time provided for completing the ratings for each of the stations?



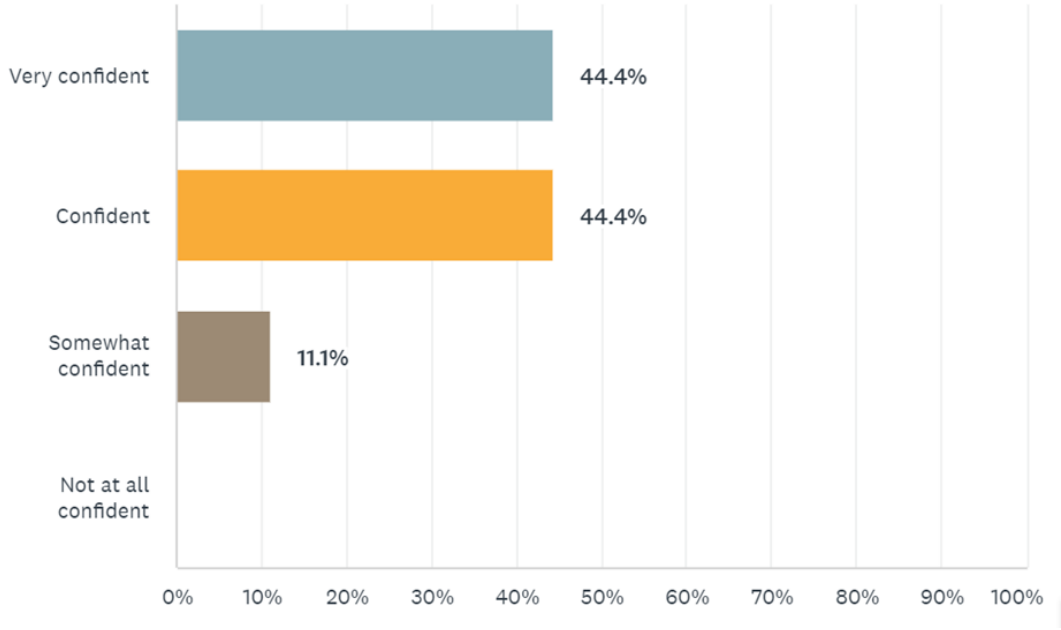
17. Overall, how did you feel about participating in group discussions conducted during the ratings process for each station?



18. How helpful was the impact data and discussion in facilitating the panel to arrive at pass scores?



19. What level of confidence do you have in the final recommended cut score for “pass”?



20. What level of confidence do you have in the final recommended cut score for “pass with superior performance”?