National Assessment Collaboration

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

Psychometrics and Assessment Services

July 2019

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 and certification decisions in the health professions. We need to assure the public that licence and certificate holders possess the required knowledge, skills and attitudes necessary for safe and effective patient care. Standard setting is a process used to define an acceptable level of performance in the competency domains targeted by an examination. The resulting conceptual standard is operationalized as a numerical pass score that is used to make classification decisions (e.g., pass/fail, grant/withhold a credential, award/deny a licence). A rigorous and valid process for standard setting should be adhered to for licensing examinations (Cizek, 2012). This report documents the processes, procedures and results of a standard- setting exercise carried out for the National Assessment Collaboration (NAC) Examination.

The NAC Examination is a one-day exam that assesses International Medical Graduates' (IMGs) readiness 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. It is a performance exam composed of a series of Objective Structured Clinical Examination (OSCE) stations (10 operational and two non-scored pilot stations), which depict various clinical scenarios including problems in medicine, pediatrics, psychiatry, surgery, obstetrics and gynecology, and preventive medicine and public health.

Each station has a Standardized Patient (SP) portray the clinical scenario and each candidate is required to interact with the SP as if with a real patient. The candidate's performance on each station is observed and evaluated by a physician examiner (PE) according to a standardized scoring instrument that includes a checklist of tasks, answer key to 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, or vice versa. 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 raw total exam score. Score comparability across test forms are established through statistical linking.

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 previous pass score for the NAC Examination was established in 2013 through a standard-setting exercise. The exam has undergone significant changes and the enhanced NAC Examination was launched in March 2019, which is different from the exam prior to 2019 in terms of test specifications, format, and scoring approach. These changes warranted a new standard-setting exercise. In addition, it is best practice to review the standard and the pass score regularly to ensure that they 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.

The NAC Examination is one of the requirements for IMGs applying for the Canadian Resident Matching Service (CaRMS). The CaRMS is a very competitive process due to a large number of applicants for a small number of residency positions for IMGs each year. Consequently, some residency programs use the NAC Examination scores to rank candidates or set their own screening cut scores to deal with high volume of applicants.

The NAC Examination is not designed for ranking candidates as it does not have score precision along a wide range of the score scale due to a limited number of stations that can be realistically administered in an exam session. Rather, it is designed to focus score precision around the cut score to facilitate accurate pass/fail decisions. In addition, the cut scores set by different programs are either norm-referenced or arbitrary in nature. The NAC Examination is a criterion-referenced exam for which passing means a candidate has demonstrated an acceptable level of knowledge, skills and attitudes and is considered as meeting the standard regardless of the performance of other candidates.

In an effort to discourage residency programs from ranking candidates or setting their own cut score, the MCC decided to use a standard-setting opportunity to establish two cut scores on the NAC Examination, one for minimally acceptable performance indicating that a candidate has demonstrated an acceptable level of knowledge/skills for entering residency and the other for highly qualified performance. The first cut score was to be used for determining pass/fail status on the exam. The purpose of the second cut score is to help programs select candidates who will most likely succeed in residency when faced with high volumes of qualified candidates, at their behest. The plan is to implement the minimally acceptable (i.e., pass/fail) cut score once approved by the NEC and to implement the highly qualified cut score only after adequate data has been gathered to assess its impact and empirical evidence has been collected to support its use.

From April 24 to 26, 2019, a panel of 21 physicians from across Canada met at the MCC's office in Ottawa to participate in a standard-setting exercise for the NAC Examination. Staff from the Psychometrics and Assessment Services (PAS) directorate, with support from staff in the Evaluation Bureau (EB), facilitated the meeting. 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 exercise that led to the recommendation of two cut scores for the NAC Examination.

2. Procedures

In this section, we present how we selected the standard-setting methods, selected and assigned panelists to two subpanels, the materials we prepared and provided to the panelists prior to and during the three-day 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 a number of considerations.

- First, the NAC Examination is a criterion-referenced exam for which a cut score should be defined as an acceptable amount of knowledge that candidates must possess or an acceptable level of performance they need to demonstrate given the intended use of the exam. Whether a candidate has achieved a certain performance level (e.g., minimally acceptable, highly qualified) is determined by comparing an individual candidate's performance to 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.
- Secondly, 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 licensure and certification examinations similar to the NAC Examination (e.g., USMLE Step 2 Clinical Skills).
- Finally, the Contrasting Groups method is relatively easier to use when setting two cut scores than the Borderline Group method we previously used for setting one cut score for the NAC Examination in 2013. Given the time constraints of a three-day meeting, we used the Contrasting Groups method so that panelists did not have to repeat two rounds for each of the two cut scores (as would be needed 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).

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 an external criterion or other

method to classify candidates into two categories (e.g., qualified vs. unqualified, masters vs. non-masters). Then the 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 score distributions of the two groups are graphed and the cut score is set at the intersection (or mid-point of the intersection zone) of the two distributions if false-positive and false-negative errors are of equal importance, or moved to the right or the left to minimize the error of greater concern (De Champlain, 2013; Downing, Tekian & Yudkowsky, 2006).

In our application of the Contrasting Groups method (as is typical in medical education field), we did not use an external criterion to classify candidates. Instead, we asked standard-setting panelists to review the score sheet of each candidate on each OSCE station on the NAC Examination, make a holistic judgment on the candidate's performance on that station, and rate it into one of three levels: 1-*unacceptable*, 2-*minimally acceptable* (at least), and 3-*highly qualified*. 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 a PE during the exam session. The score distributions of the three groups were plotted. The midpoint of the intersection zone between groups 1 and 2 was selected as the cut score for *minimally acceptable* performance and the mid-point of the intersection zone between groups 2 and 3 was selected as the cut score for *highly qualified* 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 its 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 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 whereas 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 score scale in an effort to define a cut score (see section 3.3).

The Hofstee method is usually not used as a standalone method. For our purpose, we used it to define a range of cut scores to provide a "reality check" on the first cut score (i.e., for minimally acceptable performance) 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 a panel of 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 examinee population, with respect to background and educational characteristics (De Champlain, 2013).

In July 2018, the MCC sent out an email invitation to physicians across the country to solicit interest in participating in our standard-setting exercise. This solicitation resulted in over 300 interested physicians, each of whom completed a demographic information form. The original invitation email and demographic form are included in Appendix A.

Table 1: Demographic information by standard-setting subpanel

Variable of interest	Group	Subpanel 1	Subpanel 2	Total
Gender	Male	3	4	7
	Female	7	7	14
Geographic region	West	4	3	7
	Prairies	1	2	3
	Ontario	3	4	7
	Quebec	1	1	2
	Maritimes	1	1	2
Ethnic background	Caucasian	5	6	11
	Other	5	5	10
Specialty	Family Medicine	5	5	11
	Other specialties	5	6	10
Years in practice post-residency	1-10	3	4	7
	11-30	7	6	13
	30+	0	1	1
Practice community	Urban	9	9	18
	Rural	1	2	3
Care setting	Hospital-based	6	6	12
	Community-based	4	5	9

Based on the demographic information provided, the MCC selected 21 participants (we originally selected 22 but one withdrew) and assigned them to two subpanels that were matched as closely as

possible on key demographic variables, including: (1) gender, (2) geographic region, (3) ethnic background, (4) medical speciality, (5) number of years in practice post-residency, (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 opportunity to share their perspective. Table 1 summarizes the demographic composition of the two subpanels.

2.3 Preparing materials for the standard-setting exercise

Preparing well for the standard-setting exercise is key to a smooth and successful standard-setting exercise. Preparation involved assembling materials for a training station to train panelists and allow them to practice using the Contrasting Groups and Hofstee methods and for the actual operational stations that were used in establishing the two cut scores. In addition, and perhaps the crux of any standard-setting exercise, we defined performance levels for the unacceptable, minimally acceptable and highly qualified candidate.

2.3.1 Materials for the training station

A non-counting pilot station was used as the training station for training panelists on the standard-setting procedures. Three video-taped candidate performances on the training station were selected, each demonstrating a performance that was unacceptable, minimally acceptable, or highly qualified. A separate binder of materials was prepared for each panelist for training and practice purpose. It included candidate instructions, scoring key, score sheets for the three video performances, and 75 score sheets ordered from the lowest to the highest score on the training station.

2.3.2 Materials for the operational stations

A stratified random sample of 75 candidates were selected from the March 2019 exam cohort based on their raw total exam scores to cover a wide range of scores. The score sheets of the 75 selected candidates on each station for a total of 750 score sheets for the ten operational stations were used for standard setting. Each score sheet represented a candidate's station performance profile and it included the candidate's scores on checklist items, oral questions and competency rating scales on a station provided by a PE during the exam session. A binder of materials for the ten stations was prepared for each panelist for the standard-setting exercise. For each station, it included candidate instructions, props if applicable, scoring key, score sheets for two video performances (see below), and 75 score sheets ordered form the lowest to the highest station score. The order of score sheets differed from station to station as candidate performances varied by station.

Two video-taped candidate performances on each station were prepared for the standard-setting exercise, one representing a minimally acceptable performance and another representing a highly qualified performance. The videos were selected by physician subject matter experts from the actual candidate performances recorded in two test centers (with informed consent from candidates, PEs and SPs).

2.3.3 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 two targets: one was a *minimally acceptable* candidate, the other a *highly qualified* candidate, and both were defined in terms of competence required for entry into residency training in Canada.

The performance levels for the *unacceptable, minimally acceptable* and *highly qualified* candidates were defined through a one-day focus-group meeting in advance of the standard-setting exercise by a different group of physicians who were knowledgeable about medical education, resident selection, and who were familiar with the IMG population. The performance levels were defined in the context of three competency domains: assessment/diagnosis, management and communication. These definitions are included in Appendix B.

2.4 Advance 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; (2) performance level definitions and; (3) two research papers that provided overviews of standard setting (De Champlain, 2013; Downing, Tekian & Yudknowsky, 2006).

2.5 Activities during the three-day 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, followed by two rounds of the standard-setting exercise over the remainder of the three-day meeting. 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 heavily on extensive 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 an introduction of facilitators and panelists as well as an overview of the purpose of the meeting. We specifically told panelists that their task was to recommend two cut scores, not to make final decisions, 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 emphasized how the enhanced and newly implemented NAC Examination differed from its predecessor. Next, we followed with an overview of the standard setting including 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, envisioned some *minimally acceptable* and *highly qualified* candidates, discussed their characteristics, what they knew or were capable of doing, things they might have difficulty in doing, what distinguished *minimally acceptable* from *unacceptable* candidates, what distinguished *highly qualified* from *acceptable* candidates, etc. The MCC's medical education advisor, who is also a practising physician, facilitated the discussion. The purpose was to calibrate the panelists to a common understanding and expectation of an appropriate level of performance 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 the following 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 unacceptable, minimally acceptable, and highly qualified candidate 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., 1-unacceptable, 2-minimally acceptable, 3-highly qualified), and entered their ratings in the standard-setting tool on the computer. As described in section 2.3.1, the score sheets were ordered from the lowest to the highest station score.

Through training, hands-on practice and thorough discussions, panelists developed a very good understanding of the performance level definitions, standard-setting process, and their specific tasks.

2.5.2 Standard-setting exercise

Following the training, we then proceeded to two rounds of the standard-setting exercise.

2.5.2.1 Initial round

For the initial round, we split the panelists into two subpanels and placed them in two different rooms. A psychometrician and a TDO facilitated each subpanel. For each of the ten 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 *minimally acceptable*, the other for *highly qualified* performance).

Initially, we gave panelists approximately 90 minutes per station to complete the rating task. As panelists became more familiar with the process, they were able to complete the task in approximately 40 minutes per station for the last few stations. 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 ten 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 *minimally acceptable* performance). Specifically, we asked panelists to specify the highest and lowest cut scores as well as the highest and lowest failure rates that they felt would be reasonable for the NAC Examination based on their holistic judgment of the purpose and content of the exam and the intended test-taker population.

We calculated the two cut scores (see section *2.5.2.3*) 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 March 2019 cohort. In addition, we used panelists' answers to Hofstee questions to define a range of 'acceptable' cut scores (for *minimally acceptable* performance) 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 minimally acceptable performance) and the percentage who fell below the second cut score (i.e., for highly qualified 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 minimally acceptable performance) by subpanel and full panel;
- Historical pass rates.

Afterwards, panelists discussed the results and impact data, first within subpanel and then with the full panel.

The initial round exercise provided panelists with an opportunity for realistic practice in full scale. 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. With the information learned and skills developed from the initial round, panelists were better prepared for the final round.

2.5.2.2 Final round

In the final round, we again split panelists into two subpanels and assigned each subpanel to a separate room. 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 (1-*unacceptable*, 2-*minimally acceptable*, 3-*highly qualified*) 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 count towards setting the two cut scores. 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 allowed 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 ten stations, we gave them a 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 provided panelists an opportunity to provide feedback on the standard-setting exercise by answering an online survey anonymously.

2.5.2.3 Calculation of the two cut scores

In this section, we describe how we calculate the two cut scores for *minimally acceptable* and *highly qualified* performance, respectively.

Cut score for minimally acceptable performance:

> We used each panelist's rating for each station to derive a station cut score, which was the mid-point between the maximum score of the group of candidates rated as unacceptable and the minimum score of the group of candidates rated as minimally acceptable as illustrated in Figure 1. We repeated this process for each of the ten stations. We then obtained each panelist's cut score for the total exam by taking the median of their ten 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 minimally acceptable performance.

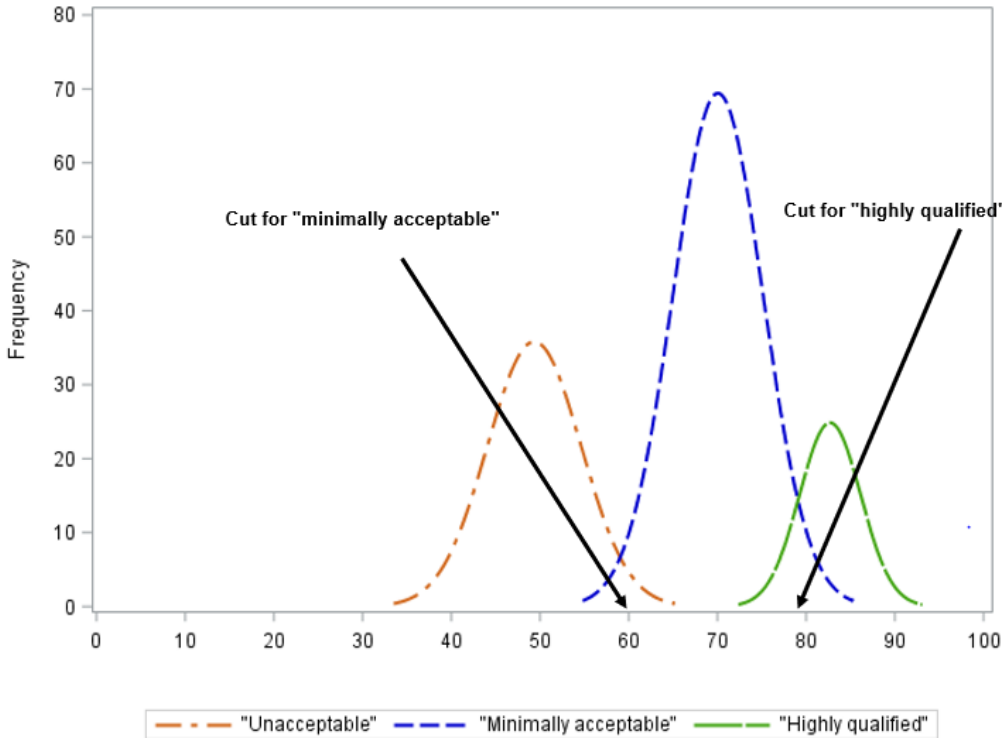


Figure 1: Illustration of cut-score calculations

Cut score for *highly qualified* performance:

The process for calculating the cut score for the *highly qualified* is the same as that for the *minimally acceptable* performance except that we used the mid-point between the maximum score of the group of candidates rated as *minimally acceptable* and the minimum score of the group of candidates rated as *highly qualified*.

3. Results

In this section, we present results of the Contrasting Groups method, generalizability analyses, impact data, and Hofstee results. We follow this with a presentation of the final recommended cut scores for NEC's review and approval. Finally, we provide results of the post-session survey.

3.1 Contrasting groups results

In Table 2, we present a summary of the computed cut scores for the *minimally acceptable* performance for each subpanel and the full panel. 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.

A summary of cut scores for the highly qualified performance is not included in this report as the plan is not to release the second cut score until further impact analyses have been completed.

Table 2: Summary of cut scores for minimally acceptable performance

		N	Median	Min.	Max.	Standard Deviation
Initial round	Subpanel 1	10	50.9	35.9	64.8	9.1
	Subpanel 2	11	48.3	17.0	61.0	13.9
	Across subpanels	2	49.6			
Final round	Subpanel 1	10	49.6	33.6	56.9	7.0
	Subpanel 2	11	49.0	27.1	63.2	9.8
	Across subpanels	2	49.3			
Final recommended cut score						49.3

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 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 in our case, and subpanel). The purpose of our analyses was to determine how much variance in the ratings provided by panelists was attributable to sources that are undesirable, such as panelists, 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 candidates into different performance levels).

We conducted a G-analysis with three facets (station, panelist and subpanel) in a candidate x station x (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 subpanel 1 and 11 in subpanel 2). We used the ratings obtained from 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.

The largest facet, not surprisingly, was the candidate-by-station interaction, which accounted for 46.7% 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 14.8% 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 facetwhich accounted for 4.7% 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 difficulty level is dependent on the 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 4.7% of total variance; station x (panelist:subpanel) explaining 2.0% of total variance and; candidate x (panelist:subpanel) accounting for 0.7% of total rating variance. Together, approximately 7.4% of the total rating variance was due to the panelist nested within the subpanel. In other words, there was some 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-by-subpanel and station-by-subpanel effects accounted for little rating variance (i.e., <=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 *minimally acceptable* performance for the two subpanels were nearly identical.

The G-coefficient and D-coefficient for the model "candidate x station x (panelist:subpanel)" were 0.75 and 0.72 respectively.

Table 3: Results of generalizability theory variance component estimates

Source	DF	SS	EMS	VCE	%Variance
Candidate	74	1447.46	19.56	0.0699	14.8%
Station	9	364.52	40.50	0.0221	4.7%
Subpanel	1	1.20	1.20	0.0000	0.0%
Candidate x station	666	3173.31	4.76	0.2201	46.7%
Candidate x subpanel	74	13.55	0.18	0.0003	0.1%
Station x subpanel	9	9.98	1.11	0.0003	0.1%
Candidate x station x subpanel	666	79.85	0.12	0.0000	0.0%
Panelist:subpanel	19	345.73	18.20	0.0220	4.7%
Candidate x (Panelist:subpanel)	1406	217.87	0.16	0.0031	0.7%
Station x (Panelist:subpanel)	171	141.45	0.83	0.0094	2.0%

DF = Degree of freedom; SS = Sum of squares; EMS = Expected mean square; VCE = Variance component estimate; % variance = Percentage of total variance

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 x station. This means that the cut scores established for this exam was dependent on the set of stations used to set the standard and would necessitate that statistical linking be implemented to ensure score comparability across exam forms (Kolen & Brennan, 2004). To address this, we will be conducting statistical linking to adjust for exam form difficulty in subsequent exam sessions so that the same cut scores can be appropriately applied over time.

3.3 Impact data

As indicated earlier, we computed the impact of cut scores using performance data from the March 2019 candidate cohort. Table 4 presents the percentage of first-time test takers and total test takers who would fall below the first cut score for each round. The overall pass rate is higher for the final round as compared to the initial round as the cut score decreased between the initial round and final round.

Table 4: Pass rates by round and candidate cohort for March 2019 exam session

	Recommended cut score	First-time test takers	All test takers
Initial round	49.6	52.8%	53.8%
Final round	49.3	53.8%	55.1%

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 *minimally acceptable*

performance. The cut score range in the final round slightly decreased from that of the initial round for both subpanels and full panels.

Table 5: Summary of Hofstee results by round and subpanels

	Statistics	Subpanel 1	Subpanel 2	Full panel
Initial round	Min. acceptable percentage cut score	52.8%	48.9%	50.9%
	Max. acceptable percentage cut score	72.0%	65.2%	68.6%
	Min. acceptable failure rate	22.5%	30.0%	26.3%
	Max. acceptable failure rate	47.5%	51.4%	49.4%
Final round	Min. acceptable percentage cut score	48.0%	47.4%	47.7%
	Max. acceptable percentage cut score	64.0%	62.0%	63.0%
	Min. acceptable failure rate	23.0%	33.6%	28.3%
	Max. acceptable failure rate	52.0%	54.1%	53.0%

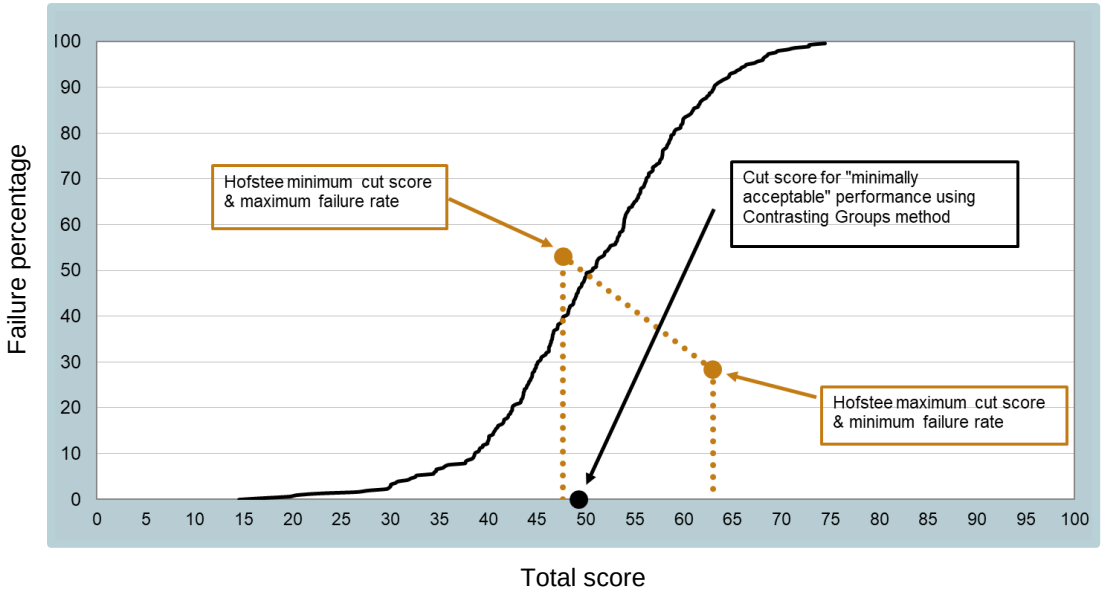


Figure 2: Hofstee results and impact

In Figure 2, the average Hoftsee answers from the full panel in the final round (as reported in Table 4) 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 from March 2019 cohort. Panelists felt that the cut score for *minimally acceptable* performance should be no lower than 47.7% and no higher than 63%. Similarly, they indicated that the failure rate should be at least 28.3% but no higher than 53%.

As indicated earlier, the Hofstee method was not our primary method for setting the standard for the NAC exam; it was used for a “reality check” of the standards set by using the Contrasting Groups method. The final cut score of 49.3 fell within the range defined using the Hofstee method. This indicates that the final cut score for *minimally acceptable* performance defined using the Contrasting Groups method was consistent with panelists’ global judgment of what the cut score and failure rate should be from political and cognitive perspectives.

3.5 Final recommended pass score

After a rigorous three-day standard-setting exercise, the panel of 21 physicians recommended a pass score of 49.3 for *minimally acceptable* performance that was subsequently brought forward to the NEC for consideration and was approved for implementation starting with the March 2019 exam session.

The implementation of the cut score for *highly qualified* performance will be considered at a later time when more data are available for assessing its impact and empirical evidence have been collected to support its use.

3.6 Post-session survey

At the conclusion of the meeting, we provided panelists an opportunity to provide feedback on the standard-setting exercise by answering an online survey anonymously. All 21 panelists responded to the survey. Full results of the survey are presented in Appendix E. The following are the highlights of the survey results.

- Central to the standard-setting exercise is the definition of the target candidate at the target level. Most panelists felt they benefited from a discussion on the *minimally acceptable* and *highly qualified* candidates and they found the discussion very helpful (66.7%), helpful (23.8%) or somewhat helpful (9.5%). The majority of respondents felt they were very clear (47.6%) or clear (38.1%) about the performance level definitions as they began the standard-setting task in the initial round and their understanding improved more in the final round (76.2% very clear and 19% clear).
- 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 90.4% of panelists felt that the amount of training was adequate. Most panelists felt that the hands-on practice was helpful (80.9%). Overall, panelists felt that the training provided was excellent (57.1%), very good (28.6%), good (9.5%) or fair (4.8%).
- 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 preliminary cut scores. Panelists found the impact data and subsequent discussions to be very helpful (71.4%), helpful (23.8%) or somewhat helpful (4.8%) in facilitating the panel to arrive at defensible cut scores.
- Finally, and most importantly, panelists indicated they were very confident (52.4%) or confident (33.3%) in the final recommended cut score for minimally acceptable performance and they indicated very confident (42.9%) or confident (42.9%) in the cut score for highly qualified 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 similar cut scores in the initial round with absolutely no influence from each other. They converged even 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 with 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, had a common understanding of the performance level definitions.
- The cut score for minimally acceptable performance defined by using the Contrasting Groups method was within the acceptable range defined by the Hofstee method based on panelists' holistic judgments. As a matter of fact, Figure 2 shows that had we used the Hofstee to define the cut score, we would have arrived at a cut score that would be very close to the cut score defined by using the Contrasting Groups method. This indicates that the criterion-referenced cut score derived using Contrasting Groups method is realistic and consistent with politically and practical considerations.
- The results of the post-session 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 evidence that the standard-setting exercise was a thorough, rigorous and valid process that meets best practice in the profession, 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 May 9, 2019, along with an overview of the standard-setting process, followed by the impact data. The NEC unanimously approved the recommended ***49.3*** as the cut score for *minimally acceptable* performance on the NAC Examination. Using the March 2019 candidate performance data, we established a new reporting scale to have a mean of 400 and a standard deviation of 25. On this new scale, the cut score was transformed to 398. This cut score will be used to determine pass/fail status on the NAC Examination and will remain in place for subsequent exam administrations.

5. References

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

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

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). San Francisco: Jossey-Bass.

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

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

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*. Thousand Oaks, CA: Sage Publications Inc.

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

July 2, 2018

Dear Doctor,

The purpose of this letter is to invite you to express interest in serving as a panelist in a standard-setting exercise for the National Assessment Collaboration (NAC).

The NAC exam is an Objective Structured Clinical Examination (OSCE) that assesses the readiness of International Medical Graduates (IMGs) for entry into Canadian residency programs. The NAC exam is required for IMGs to apply for a Canadian residency program position.

A new and enhanced NAC exam will be implemented in 2019. To set the performance standard for the new NAC exam, the Medical Council of Canada (MCC) will conduct a standard-setting exercise from April 24 to 26, 2019. The purpose of the exercise is to recommend a new pass score for candidates taking the NAC exam beginning in 2019.

We hope that you will consider volunteering to participate in our panel as your clinical expertise and practice experience are vital to the success of this standard-setting exercise. We are sending out this notice to solicit volunteers from which we will assemble the panel to ensure that the diversity of demographics, medical specialties, clinical practice experiences and contexts across Canada are well represented.

Individuals who contributed to test development or scoring processes for the NAC exam in the past few years will not be selected as panelists for the standard-setting exercise; 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 from April 24 to 26, 2019 at the MCC head office in Ottawa. Panelists will be guided through a set of procedures to evaluate examination materials to arrive at a recommended pass score. An honorarium of $600 per day (full 3-day meeting) plus reasonable travel and accommodation expenses will be provided. Panelists will also be issued a certificate of participation which they can use to apply for CPD credits.

We hope that you will be interested in participating. Should you be, we ask that you complete the demographic information survey by October 31, 2018, and tentatively reserve the standard-setting dates in your calendar. Your participation will be confirmed by November 15, 2018. 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
Psychometrics and Assessment Services, MCC
research@mcc.ca

Associate Director

mcc.ca
physiciansapply.ca
inscriptionmed.ca

Demographic Information Survey for the National Assessment Collaboration (NAC) Examination

The information requested below is being collected to help the Medical Council of Canada (MCC) select a representative, pan-Canadian panel to recommend a pass score on the National Assessment Collaboration (NAC) Examination. The standard- setting exercise will take place from April 24 to 26, 2019.

Completed surveys must be submitted by **October 31, 2018**. Should you have any questions, please contact us at *research@mcc.ca*.

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

---

1. Do you have a Licentiate of the Medical Council of Canada (LMCC)?
   (Please use a ✓ or simply highlight your choice.)
   - ○ No
   - ○ Yes (please provide your LMCC number): _____

2. 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

3. Do you have an active, unrestricted licence to practise with a Medical Regulatory Authority (MRA) in Canada?
   - ○ No
   - ○ Yes (please specify which province/territory): _____

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

5. Region 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

6. First language:

- ○ English
- ○ French
- ○ Other (please specify): _____

7. Primary language of your medical practice:

- ○ English
- ○ French
- ○ Other (please specify): _____

8. Gender identity:

- ○ Female
- ○ Male
- ○ Prefer to self-describe: _____

9. I identify my ethnicity as:

- ○ Caucasian
- ○ Indigenous
- ○ Other group (please specify): _____

10. Medical specialty:

- ○ Pediatrics
- ○ Internal Medicine
- ○ Psychiatry
- ○ Obstetrics and Gynecology
- ○ Surgery
- ○ Family Medicine
- ○ Other (please specify): _____

12. Type of community in which you primarily work:

- ○ Urban
- ○ Rural

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

- ○ Hospital-based setting
- ○ Community-based setting

14. 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

15. Have you practiced within the last three years in Canada?

- ○ Yes
- ○ No (please specify): _____

16. Have you ever participated in a NAC Examination 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.

17. Have you ever been an examiner for the MCCQE Part II or the NAC Examination? Please select all that apply:

- ○ I have been a Medical Council of Canada Qualifying Examination (MCCQE) Part II examiner
- ○ I have been a National Assessment Collaboration (NAC) Examination examiner
- ○ I have not done either
- ○ **NOTE:** Being an MCC examiner is not a requirement to participate in the standard-setting exercise.

18. Have you participated in a third-party candidate test preparatory course (i.e., not offered by the MCC) in preparation for the NAC Examination or the MCCQE Part II within the last three years?

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

19. Have you had experience supervising students/residents?

- ○ No
- ○ Yes

20. How recently have you supervised students/residents?

- ○ Within the past 1-5 years
- ○ Within the past 6-10 years
- ○ Within the past 11-20 years
- ○ More than 20 years ago

21. What number of students/residents do you typically supervise in a given year? _____

22. How many years of experience do you have supervising Canadian medical graduates (CMGs)?

- ○ 1-2 years
- ○ 3-5 years
- ○ 6-10 years
- ○ 11-20 years
- ○ 21-30 years
- ○ More than 30 years
- ○ I have no experience supervising CMGs

23. How many years of experience do you have supervising International Medical Graduates IMGs)?

- ○ 1-2 years
- ○ 3-5 years
- ○ 6-10 years
- ○ 11-20 years
- ○ 21-30 years
- ○ More than 30 years
- ○ I have no experience supervising IMGs

Thank you for taking the time to complete this survey.
Should you have any questions, please contact us at research@mcc.ca.

Appendix B: Standard-setting meeting agenda

NAC Standard-setting exercise
Wednesday, April 24 to Friday, April 26, 2019
Thomas Roddick & Maude Abbott Meeting Room

Agenda

DAY 1: WEDNESDAY, APRIL 24

TIME	ITEM
07:45 a.m.	Breakfast
08:00 a.m.	Welcome and introductions
08:15 a.m.	Security video
08:20 a.m.	Review agenda and objectives
08:30 a.m.	Overview of the NAC exam
08:55 a.m.	Overview of standard setting
09:20 a.m.	Discussion on performance level definitions
10:05 a.m.	Break
10:20 a.m.	Training and practice
11:45 a.m.	Lunch
12:30 p.m.	Station T01 (Initial round)
14:15 p.m.	Break
14:25 p.m.	Station T02 (Initial round)
16:00 p.m.	Station T03 (Initial round)
17:25 p.m.	Wrap-up of day 1

DAY 2: THURSDAY, APRIL 25

TIME	ITEM
07:45 a.m.	Breakfast
08:00 a.m.	Stations T04/T06 (Initial round)
10:20 a.m.	Break
10:30 a.m.	Station T07 (Initial round)
11:40 a.m.	Lunch
12:25 p.m.	Station T08/T10 (Initial round)
14:35 p.m.	Break
14:45 p.m.	Stations T11/T12 (Initial round)
16:55 p.m.	Hofstee method
17:05 p.m.	Wrap-up of day 2

DAY 3: FRIDAY, APRIL 26

TIME	ITEM
07:45 a.m.	Breakfast
08:00 a.m.	Present initial round results, impact data and discussion
08:45 a.m.	Stations T01/T02 (Final round)
09:55 a.m.	Break
10:05 a.m.	Stations T03/T04/T06 (Final round)
11:50 a.m.	Lunch
12:35 p.m.	Stations T07/T08/T10/T11/T12 (Final round)
15:15 p.m.	Hofstee method
15:25 p.m.	Break and tour
16:05 p.m.	Present final round results and impact data
16:20 p.m.	Post-session survey
16:30 p.m.	Wrap-up of day 3

Appendix C: Performance level definitions

The candidate's lowest deficiency in any of Assessment and diagnosis, Management or Communication competencies is how they are categorized overall.

For example, if a candidate is Minimally acceptable in Assessment and diagnosis and in Management, but they are Unacceptable in Communication, they are categorized as Unacceptable overall.

	UNACCEPTABLE	MINIMALLY ACCEPTABLE	HIGHLY QUALIFIED
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 highly qualified to enter residency training.** The candidate consistently 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 consistently 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 has the expected skill to consistently develop a clear definition of the patient's problem and 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 consistently has appropriate management strategies for common, acute and emergent illnesses. Their management plan is 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 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? _____

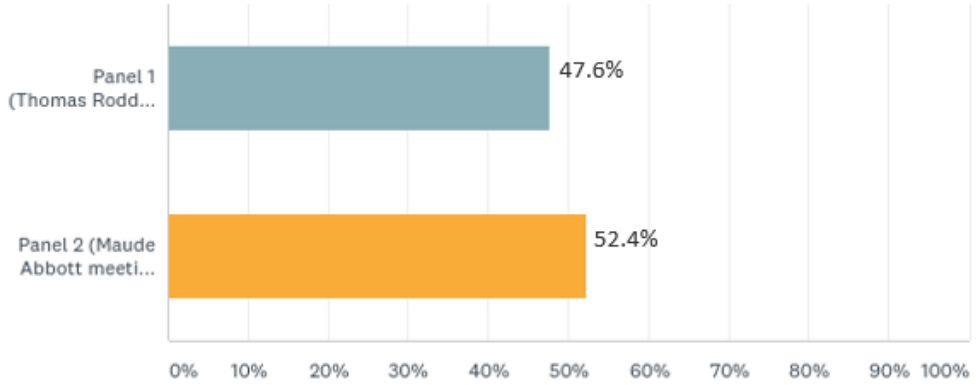
Round: Final

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? _____

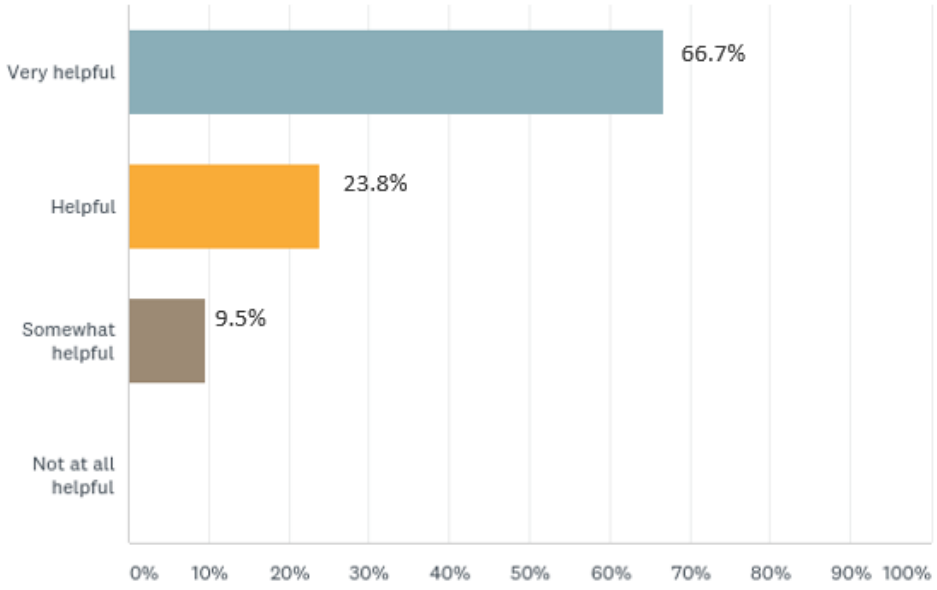
Your collective answers to the four questions will be used to define a range of acceptable pass scores that will be used to check the reasonableness of the cut score defined using the Contrasting Groups method.

Appendix E:
Summary of responses to post-meeting survey

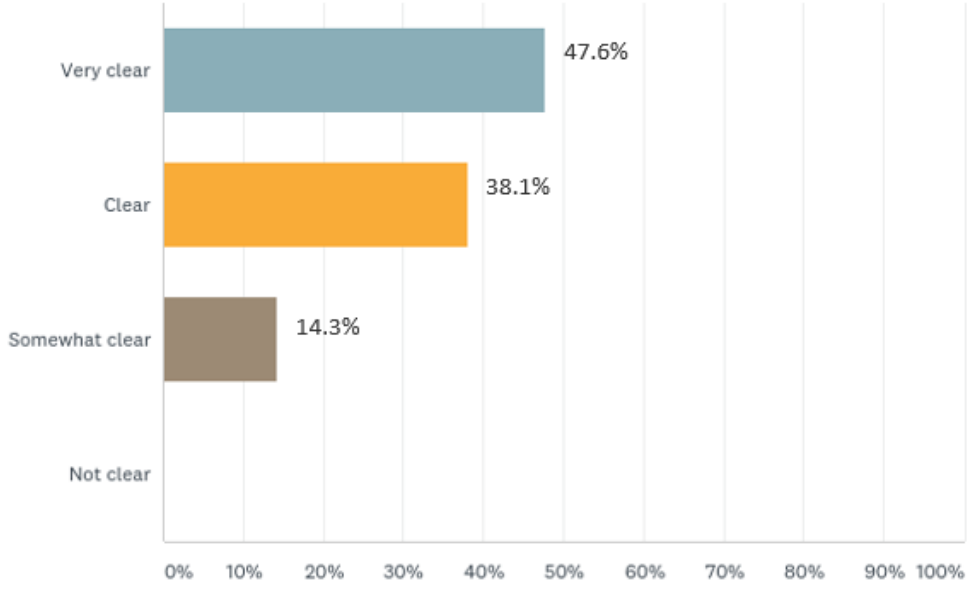
1. Which room were you in for the standard-setting exercise?



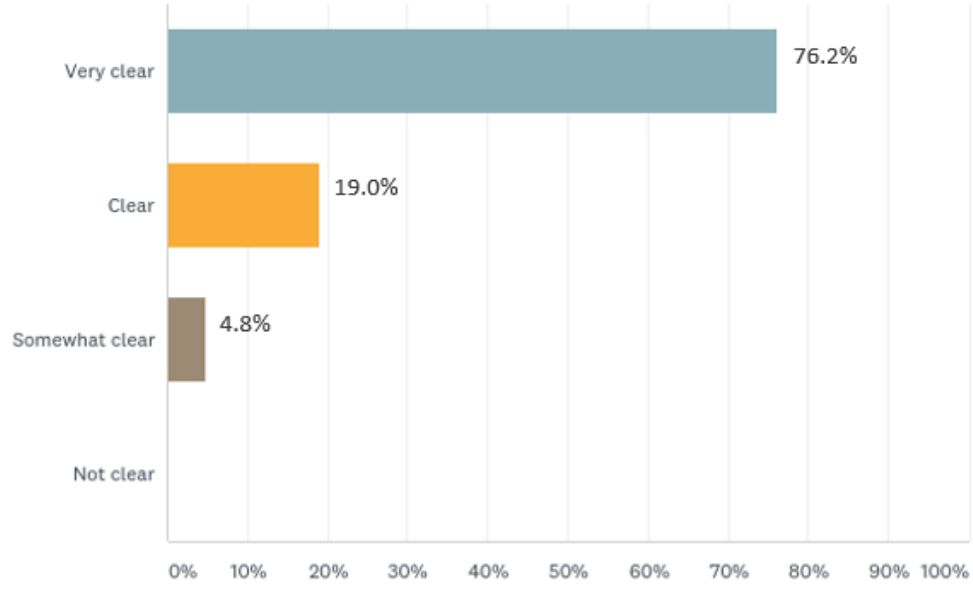
2. During the training on Day 1, how helpful was the discussion on the “minimally acceptable candidate” and the “highly qualified candidate” for the NAC exam?



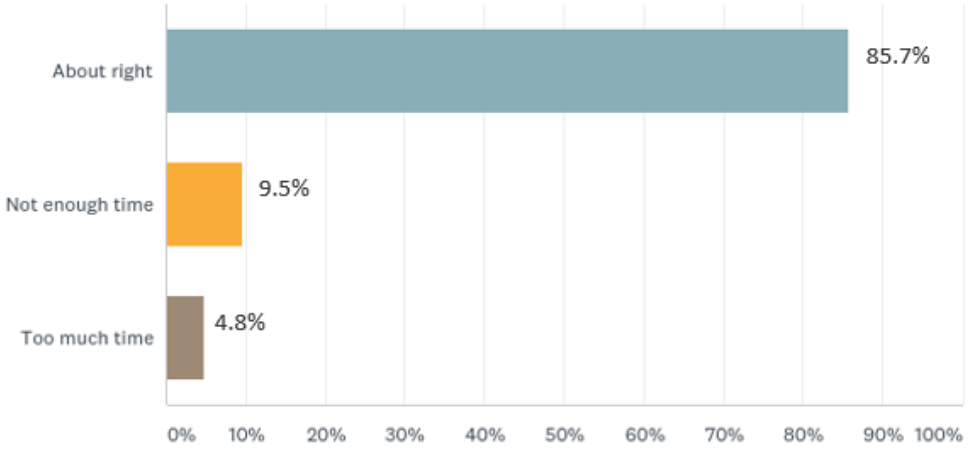
3. Following the training on Day 1, how clear was your understanding of the descriptions of the "minimally acceptable candidate" and the "highly qualified candidate" for the NAC exam as you began the task of setting cut scores in the initial round?



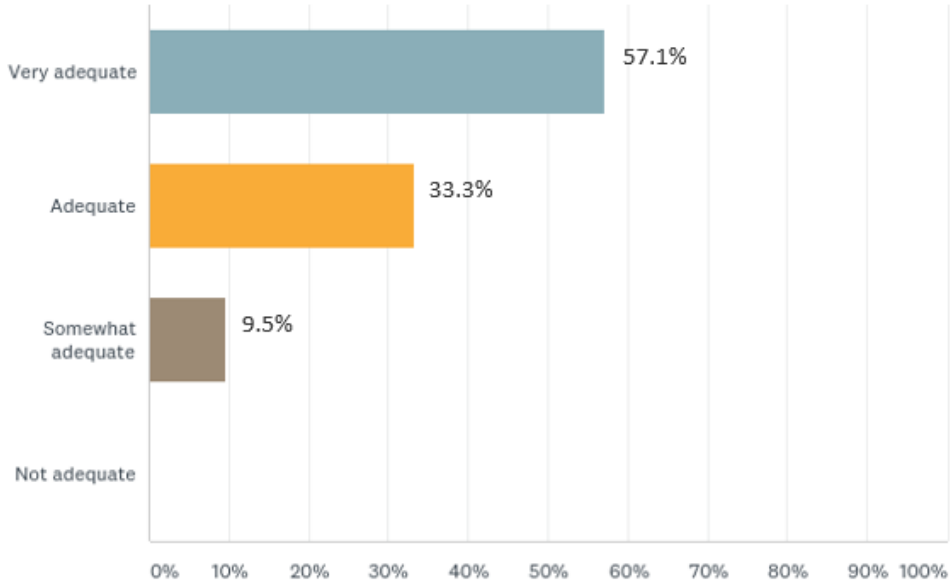
4. On Day 3, how clear was your understanding of the descriptions of the "minimally acceptable candidate" and the "highly qualified candidate" for the NAC exam as you began the task of setting cut scores in the final round?



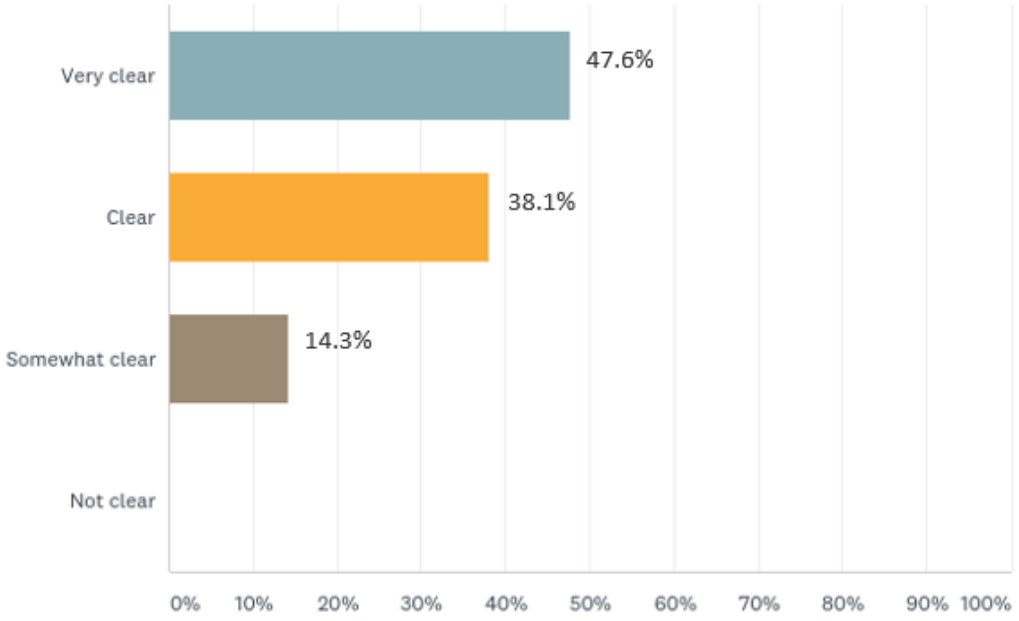
5. How would you judge the length of time spent introducing and discussing the definitions of the "minimally acceptable candidate" and the "highly qualified candidate" (approximately 45 minutes)?



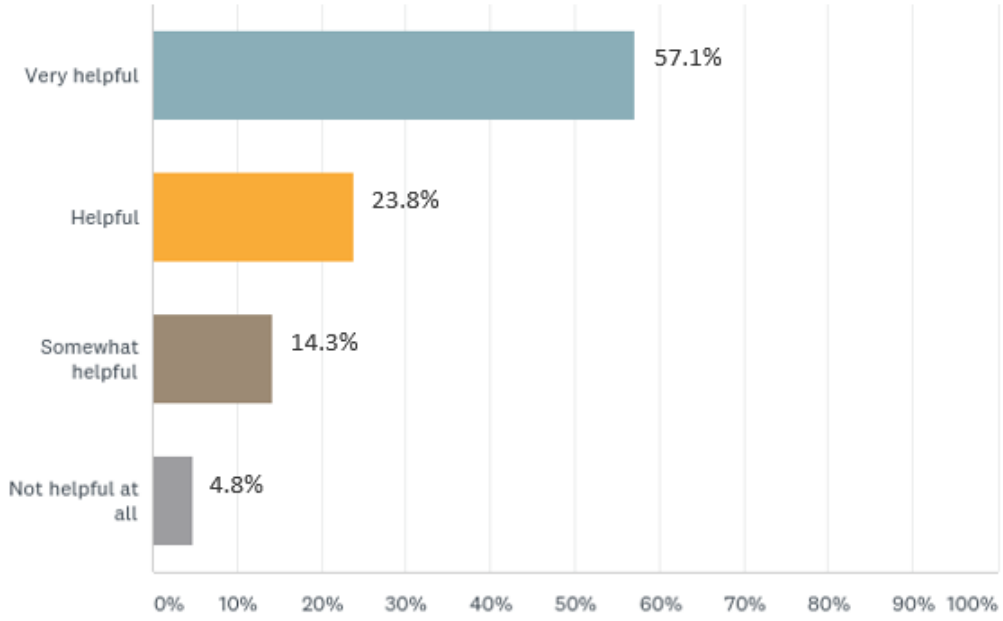
6. What is your impression of the amount of training you received on setting cut scores?



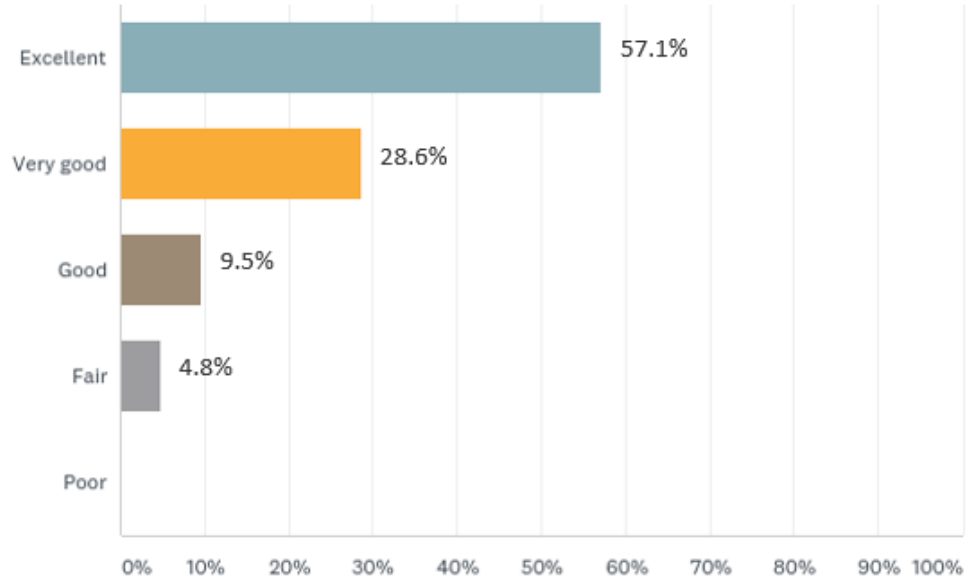
7. How clear was the information provided regarding the scoring procedures for the NAC exam?



8. How helpful was the practice session for using the electronic rating tool?



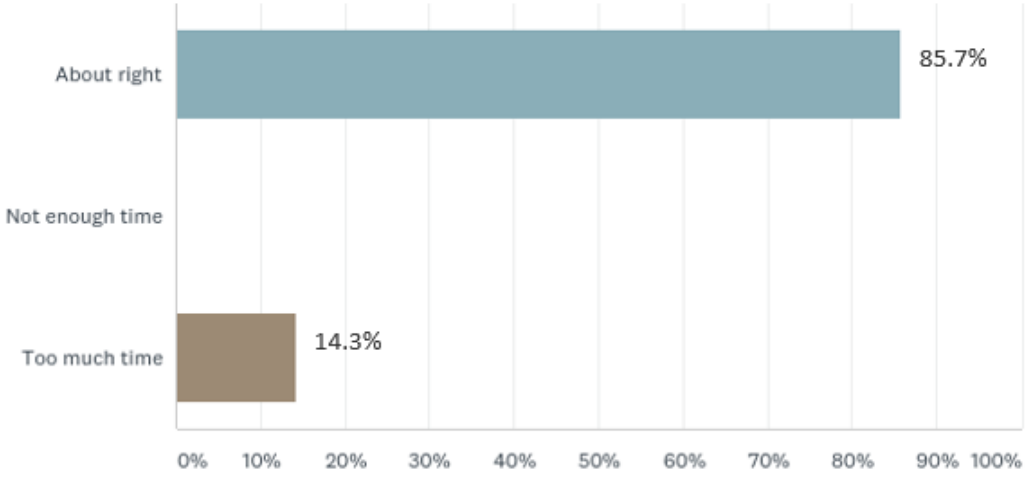
9. What is your overall evaluation of the training provided for setting cut scores for the NAC exam?



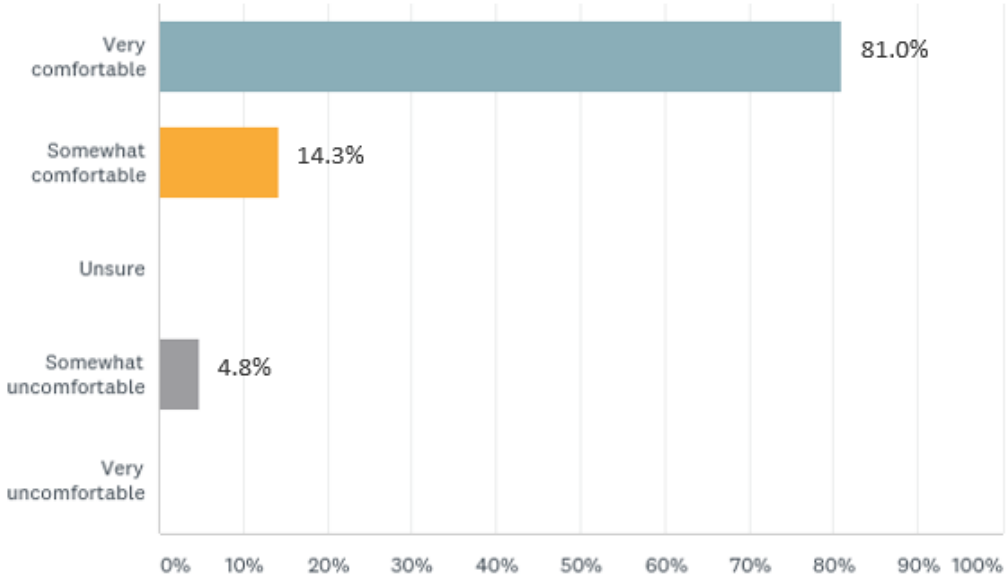
10. What factors influenced the ratings (i.e., 1, 2, 3) you made based on candidate score sheets on the NAC exam? (Select all that apply)

ANSWER CHOICES	RESPONSES	
Descriptions of the "minimally acceptable candidate" and the "highly qualified candidate"	90.5%	19
My perception of the difficulty of each station	71.4%	15
Candidate's score profiles (on checklist items, oral questions, and rating scale items)	76.2%	16
Candidate station scores	57.1%	12
The impact data provided before the final round	52.4%	11
Panelist discussions	76.2%	16
My experience in the field	52.4%	11
My experience with students/residents in the field	76.2%	16
Knowledge and skills measured by each station	71.4%	15
Other (please specify):	19.0%	4
Total Respondents: 21		

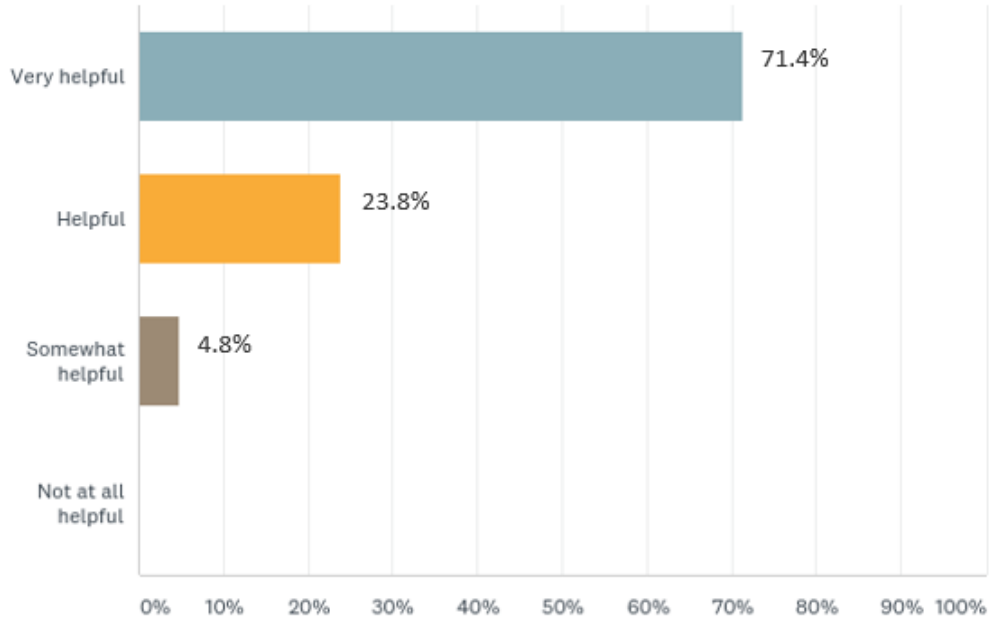
11. How would you judge the length of time provided for completing the ratings for each of the stations?



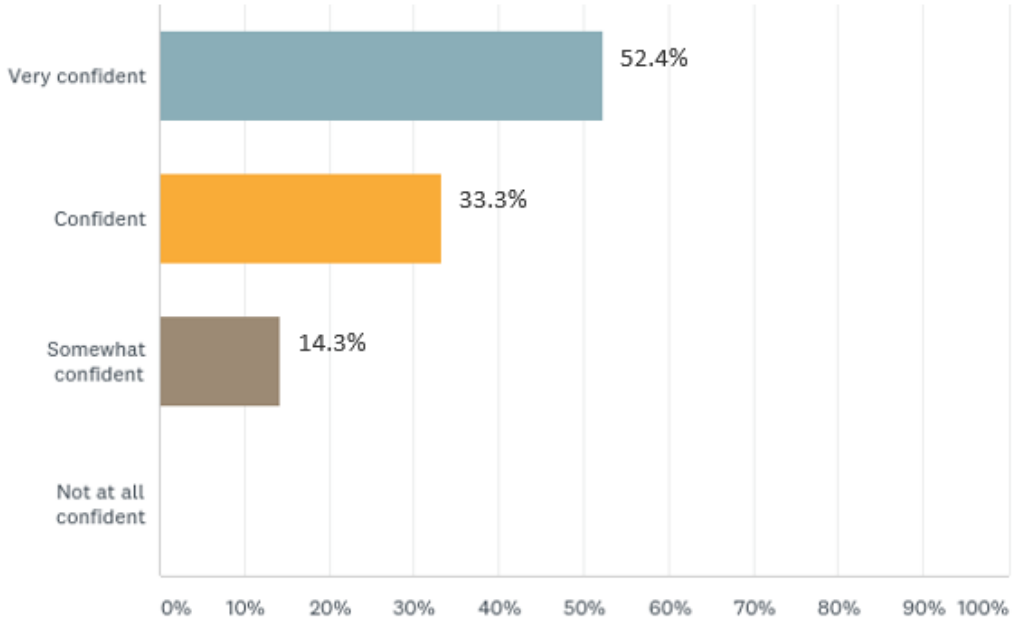
12. Overall, how did you feel about participating in group discussions throughout the standard-setting exercise?



13. How helpful was the impact data and discussions in facilitating the panel to arrive at cut scores?



14. What level of confidence do you have in the final recommended cut score for the “minimally acceptable candidate”?



15. What level of confidence do you have in the final recommended cut score for the “highly qualified candidate”?