National Assessment Collaboration Standard Setting Study Report

Prepared for the Medical Council of Canada while interning Summer 2012

Louise M Bahry, University of Massachusetts Amherst
Ron K. Hambleton, University of Massachusetts Amherst
Andrea J Gotzmann, Medical Council of Canada
Andre De Champlain, Medical Council of Canada
Marguerite Roy, Medical Council of Canada

Acknowledgements

I wish to extend a thank you to the Medical Council of Canada for providing me the opportunity to work on this project as well as all the staff who made time to meet with me and discuss the NAC OSCE exam. In particular, thank you to Andrea Gotzmann, Andre De Champlain and Marguerite Roy for their support throughout the summer. Finally, a special thank you to Dr. Ron Hambleton for his guidance and advice.

Table of Contents

Executive Summary

The purpose of this report is to propose a study to provide guidance on the selection of a standard setting method for use with the National Assessment Collaboration (NAC) Objective Structured Clinical Examination (OSCE) administered by the Medical Council of Canada (MCC). Currently, cut-scores for the NAC are set with each administration of the exam, separately for each OSCE station as well as the therapeutics portion. Changes coming into effect in the spring of 2013 will allow for a new method of setting cut-scores across administrations and a movement to set a cut-score at the total test score level is under consideration. The proposed standard setting study contained in this report is to be carried out in early 2013. The NAC OSCE examination is for International Medical Graduates (IMGs) that wish to apply to residency programs in Canada. The NAC OSCE exam consists of two sections: an objective structured clinical exam (OSCE) component and a written therapeutics component.

Hambleton & Pitoniak (2006) suggest four main considerations in choosing an appropriate method. First, the item formats comprising the assessment should be considered. There are methods to be considered when the assessment consists of multiple choice items only (Angoff, 1971; Lewis, Mitzel, Mercado, & Schulz, 2012), performance assessments (Cizek & Bunch, 2007) and mixed item assessments or score profiles (Hambleton, Jaeger, Plake, & Mills, 2000; Kingston & Tiemann, 2012). The format of the NAC OSCE exam does not permit the use of the test-centred methods generally used with educational assessments, particularly due to the preference of setting a cut-score at the total test score level. The most appropriate holistic or compromise methods for use with the NAC are the borderline or contrasting groups methods, judgemental policy capturing, analytical method, and the body of work method. These methods are outlined in detail in the body of the report.

Second, time to complete the standard setting exercise may be limited. As such, it is important to keep this in mind and align the time needed to implement a method with the needs and resources of the testing program. While there are not stringent constraints places on the time allotted to carry out tasks the standard setting in the case of the NAC, those methods requiring an inordinate amount of time for training or carrying out such as the judgmental policy capturing method and the analytic method may not be appropriate.

Third, programs that have had previous experience with a particular method should include the method in the review because field testing a new method is both expensive and time consuming. Currently, MCC will be using the Hofstee method with another exam, and because of this it may be of interest to compare this method to another method for the purposes of this study. Finally, evidence and/or perceptions of the validity of a method should be evaluated. When taking perceptions of the validity of the methods outlined into consideration, it appears the most commonly used methods in medical licensure currently are the borderline and contrasting groups methods (De Champlain, 2004). Because these appear most commonly in the literature, it may be advantageous to include one of the methods in the study for comparison.

The comparison of different methods has been studied, and results suggest that often different methods will produce different results (Cizek & Bunch, 2007). It is recommended that a method be chosen because it matches the format and purpose of the assessment. However, given there are several methods that appear to match the needs for the NAC OSCE exam and the exploratory nature of the proposed study, a comparison of two or more methods seems both possible and appropriate.

All of the methods discussed previously involve a panel of experts. The selection of a sufficiently large, representative panel is another major consideration of a standard setting study.

In order to attempt to provide some evidence of generalizability of the cut-score, Hambleton, Pitoniak, & Copella (2012) suggest selecting double the required panellists and having two parallel panels carry out the standard setting procedures. This design would allow for an investigation of the generalizability of the cut-scores set in the study.

The final section outlines the practical considerations in carrying out the proposed study. First, three different study designs are proposed each with possible positive and negative consequences. Second, a discussion of required training for the study is presented, including contingencies dependent on the design carried out. Extensive training of panelists before the standard setting will help them to focus on the true purpose of the exam instead of extraneous details that may unduly influence their ratings. Training provides an opportunity for panelists to ask questions and have initial discussions allowing orientation to the task at hand. When panelists have a clear and deep understanding of their role in the meeting, it can also lessen the time needed to carry out the standard setting session.

Following from the review of the literature, the three methods recommended for comparison are discussed including details regarding the minimal number of cases to be reviewed and the number of rounds to be carried out. Finally, the document concludes with a list of the materials recommended for the standard setting study to compare the three methods. Samples of some of the materials are included in the appendices.

Section 1. Introduction

The purpose of this report is to propose a study that will guide the selection of a standard setting method for use with the National Assessment Collaboration (NAC) Objective Structured Clinical Examination (OSCE) administered by the Medical Council of Canada (MCC). Standard setting is a process by which one or more cut-scores, passing scores, minimum achievement levels, etc., are chosen on an assessment which determine specific points on the score scale that classify examinees into particular groups of interest (e.g., pass/fail, proficient/advanced; (Boulet, De Champlain, & McKinley, 2003; Cizek, 2012b; Cizek, Bunch, & Koons, 2004). With credentialing exams like the NAC, normally only a single cut score is needed to distinguish passing from failing candidates.

Currently, cut-scores for the NAC are set with each administration of the exam, separately for each OSCE station as well as the therapeutics portion. Changes coming into effect in the spring of 2013 will allow for a new method of setting cut-scores across administrations and a movement to set a cut-score at the total test score level is under consideration. The proposed standard setting study contained in this report, subject to review and modification by the staff, the TAC and consultants, will is to be carried out in early 2013.

This report has been organized into three distinct sections. First, an introduction and full description of the NAC OSCE examination will be provided to focus the discussion of literature and methods that follow. Next, a section on the current literature is presented that guides best practice in selecting a standard setting method, panelists, and validation considerations. The third and final section of the report presents a collection of methods for possible use with the NAC OSCE exam standard setting study, a discussion of the importance of training, three possible designs in structuring the study, and a list of the materials required for training and carrying out the standard setting.

The National Assessment Collaboration (NAC) Examination

The NAC OSCE examination for International Medical Graduates (IMGs) that wish to apply to residency programs in Canada. The NAC OSCE exam consists of two sections: an objective structured clinical exam (OSCE) component and a written therapeutics component. The OSCE section of the exam consists of 10 scored and 2 pilot (2 different stations per site) OSCE stations approximately 11 minutes in length. OSCE stations require candidates to interact with standardized patients while being evaluated by a physician examiner on up to nine competencies: history taking, physical examination, organization skills, communication skills, language fluency, differential diagnosis, data interpretation, investigations, and management.

Individual stations measure a sample of these nine competencies with a minimum of five in a single OSCE station. In each of the 10 scored stations, competencies are assessed on a rating scale of 1-5 (1 being unacceptable, borderline unacceptable, borderline acceptable, acceptable, and 5 being above expected) and a global rating is provided by the examiner on the same 1-5 scale for each station. The cut-score for the OSCE component is calculated by station by taking average of the “borderline candidate” scores for each station, and then taking the average of the 10 station cut-scores.

For 2012, the therapeutics component of the NAC consists of 12 short answer and 18 multiple choice questions (MCQs) 12 operational and 6 pilot (6 different MCQs per site). In 2013, the therapeutics component will change to consist of 24 multiple choice questions only. The 24 counting items are aligned with 14-18 clinical scenarios and cross four distinct sections: pharmacotherapy, adverse effects, disease prevention, and health promotion. Two physician examiners score according to a pre-defined rubric the 12 short answer questions. Two different physician examiners provide a global rating for the entire therapeutics component on the same 1-5 rating scale used in the OSCE component of the exam. The TPx component cut-score is calculated by taking the average of the two physician examiner ratings and those ratings that are between 2.51 and 3.49 are in the “borderline” group and the average for those candidate’s total TPx score is the cut-score for that component. The final cut-score is calculated by weighting the OSCE component 75% and the TPx component 25%.

Section 2. Literature Review

The purpose of this section is to outline the nine steps recommended by Hambleton, Pitoniak, and Copella (2012) in carrying out a defensible standard setting and incorporate relevant literature to inform the proposed study. Step one includes the selection of an appropriate method and the preparation for the meeting of the standard setting panel. Step two involves the selection of the panel and design of the study and three includes the preparation of performance level descriptors for use by the panelists. Step four includes the task of properly training panelists to carry out the standard setting. Steps five, six and seven include the compilation of ratings provided by panelists in the standard setting, providing feedback to panelists to facilitate discussion, and a second compilation of ratings to obtain the actual cut-score. Step eight involves conducting an evaluation of the standard setting process and nine involved compilation of all documentation and validity evidence. This section will focus primarily on providing direction regarding the selection of methods to be investigated as well as considerations regarding the selection of panelists.

Considerations When Selecting a Method

While there are many methods by which one can go about setting cut-scores, and it is recognized that there is much overlap in the methods (Cizek, 2012c), they generally fall into one of two categories: test-centred and examinee-centred (Jaeger, 1989). Generally speaking, these categories refer to the type of information that the panel of experts refers to in making their decision: test content or test takers, respectively.

Test-centred methods of setting cut-scores involve an expert panel reviewing test items and scoring rubrics to provide classifications or probabilities of correct examinee responses. While these types of methods (Angoff, 1971; Lewis, Mitzel, Mercado, & Schulz, 2012; Nedelsky, 1954) are used widely in large-scale assessment and have been shown to provide reliable and valid cut-scores (McKinley, Boulet, & Hambleton, 2005), assume an underlying unidimensional structure which cannot be assumed in the case of the OSCE portion of the NAC OSCE examination or when viewing the entire examination as whole.

Conversely, examinee-centred methods are focused on the expert panelists making more global judgements of candidate performances. These types of methods (Berk, 1976; Hambleton & Plake, 1995; Kingston & Tiemann, 2012; Livingstone & Zieky, 1982; McKinley et al., 2005) are more commonly seen in the medical education literature and have been used extensively in setting cut-scores for OSCEs (Boulet et al., 2003; Boursicot, Roberts, & Pell, 2007; Kramer et al., 2003; McKinley et al., 2005).

Hambleton & Pitoniak (2006) suggest four main considerations in choosing an appropriate method. First, the item formats comprising the assessment should be considered. There are methods to be considered when the assessment consists of multiple choice items only (Angoff, 1971; Lewis et al., 2012), performance assessments (Cizek & Bunch, 2007) and mixed item assessments or score profiles (Hambleton, Jaeger, Plake, & Mills, 2000; Kingston & Tiemann, 2012). Second, as time can be limited when completing a standard setting, it is important to keep this in mind and align the time needed to implement a method with the needs and resources of the testing program. Third, if a program has previous experience with a particular method, it should be considered as field testing a new method is both expensive and time consuming. And finally, as is being considered here, evidence and/or perceptions of the validity of a method should be evaluated.

For the purposes of setting cut-scores on the NAC OSCE exam, the test format necessarily dictates the use of holistic or compromise methods that consider the candidate’s complete score profile. What follows is a description of several of the most common and appropriate examinee-centred standard setting methods used with performance assessments in education and licensure.

Review of Relevant Methods

Contrasting and Borderline Groups

The contrasting (Berk, 1976) and borderline groups (Livingstone & Zieky, 1982) standard setting methods are two of the most commonly used methods for performance assessments (Boursicot et al., 2007; Dauphinee, Blackmore, Smee, Rothman, & Reznick, 1997; Humphrey-Murto & MacFayden, 2002; Kramer et al., 2003; Smee & Blackmore, 2008). In both cases, panelists are required to classify candidates in two or more categories.

In the contrasting groups method, panelists classify candidates in only one of two categories (e.g., pass or fail, proficient or needs improvement, etc.) whereas the borderline groups method requires classification into three categories, pass, fail, and borderline (although often in practice, candidates are asked to identify only the borderline papers). Choosing a cut-score in the either method requires operational or field test data be available. Test scores for each of the categories used in the procedure are plotted and the cut-score resides in the middle of the 'borderline' distribution for the borderline groups method and at the point where the two categories are most clearly differentiated for the contrasting groups method.

Both methods require a large sample of candidate performances to produce valid and reliable cut-scores and are not recommended for smaller scale testing programs (McKinley et al., 2005). A sample that is too small in the 'borderline' group may produce unstable estimates of the cut-scores (Mills, 1995). The recommended minimum cases for each method are noted in section 3. When using the contrasting groups method, it is possible to encounter the problem wherein the two score distributions are not clearly separated and so the placement of a cut-score is not obvious (Hambleton & Pitoniak, 2006). For both methods, it is integral to find a group of panelists familiar with the candidates and to train them appropriately to do the task.

Body of Work Method

The body of work method (Kingston & Tiemann, 2012) has become one of the most popular holistic methods developed in 1993 (Cizek & Bunch, 2007) and is has been used in several state-wide assessment programs(Hambleton & Pitoniak, 2006). The method involves two rounds of decisions by panelists called rangefinding and pinpointing, respectively. To begin, representative work samples are ordered by total score permitting the panelists to move more quickly through the process. During the rangefinding round, panelists inspect each performance sample and provide a rating placing it into a performance category. Panelists are permitted to discuss ratings with one another and to change their initial ratings as a result of the discussions.

Between rangefinding and pinpointing, those performance samples wherein there is nearly perfect agreement are removed from the samples and new samples are introduced with total scores close to those that produced variability in panelist classification. The samples resulting in near perfect agreement on classification are likely not representative of samples surrounding the cut score range. However, those performance samples that did produce variability in panelist classifications may indicate the presence of a cut-score(Cizek & Bunch, 2007). That is, those performance samples that were not clearly placed into a performance category by panelists likely lie in the “borderline” group. The pinpointing round task is the same as rangefinding, wherein the panelists provide ratings of the performance samples; however, pinpointing is done individually.

The advantage of the body of work method is the intuitive appeal of the task. Making judgements based upon work or performance samples is something panelists likely do on a regular basis (Cizek & Bunch, 2007). However, as with the borderline or contrasting groups methods, body of work requires real data in the form of total test scores as well as a large number of performance samples to use during the standard setting. A study by Kingston, Kahl, Sweeney, & Bay (2001) suggests that this method may produce somewhat inflated cut-scores than other methods. The authors suggest that providing the panelists with more information regarding the assessment may address this issue.

Judgemental Policy Capturing Method

The Judgemental Policy Capturing Method (Jaeger, 1995) requires panelists to make judgements about candidate score profiles by placing them into categories that reference the performance standards (Hambleton et al., 2000) after extensive training all possible scores a candidate could earn on each item. The initial ratings provided are analysed and mathematical models for each examinee are created. Once the models are created, panelists specify a minimum performance score required to pass the exam, and these values are collected and summarized before becoming the cut-score. This method allows panelists to provide a truly holistic judgement regarding the candidate's performance.

The policy capturing method is best used when there are few stations/items, and when scores on the assessment can take on more than two values. Additionally, the method performs well when used with a multidimensional assessment and only one cut-score is required (Hambleton et al., 2000). Jaeger (1995b) found this method produced somewhat inflated results as opposed to other methods but noted that a more iterative process may be helpful (Jaeger, 1995a). However, this method requires a large number of score profiles to be rated due to the modeling procedures used, and the training required to carry out the procedure is quite extensive and time consuming.

Analytical Method

In a standard setting meeting using the analytical method (Hambleton et al., 2000) panelists review a range of performance samples on all sections of an assessment individually. That is, if there is an essay and a multiple choice portion of an assessment these portions would be reviewed and rated separately. Once the initial ratings of the first section are completed the panelists review ratings as a group and are given an opportunity to make changes to their ratings. Next, panelists rate the second portion of the assessment. Cut-scores for each section are summed to produce a cut-score for the overall assessment. While this method is straightforward, it does require many candidate performance samples, preferably near the cut-score to enhance reliability. As with many holistic methods, the preparation for the analytical method may be time consuming as well as the standard setting exercise.

Hofstee Method

The Hofstee method (Hofstee, 1983) was developed as a “compromise” method to take into account political considerations regarding pass/fail rates for a testing program. This method allows for a look at the outcomes of choosing a particular cut-score. In that way, the Hofstee method can provide some validity evidence relates to testing consequences. In order to utilize the method, there must be previous test score data available, and panelists are asked four questions:

1. What is the highest percent correct cut-score that would be acceptable, even if every examinee receives that score?
2. What is the lowest percent correct cut-score that would be acceptable, even if no examinee receives that score?
3. What is the maximum acceptable failure rate?
4. What is the minimum acceptable failure rate?

The values indicated by panelists in response to these questions provide the coordinates for a Hofstee plot. The plot aids in the determination of whether or not the results of the cut-scores are consistent with the political considerations of cut-scores and pass/fail rates (De Champlain, 2004).

The Hofstee method is a fairly simple method to implement and takes little time to carry out. However, the method was developed for use where only pass/fail decisions were necessary and has not been extended to use where there is a need for multiple cut-scores. Also, in order to use the Hofstee method, a prior distribution of scores is necessary to construct the Hofstee plot (Cizek & Bunch, 2007).

Selection of Methods

As noted above, Hambleton & Pitoniak (2006) suggest there are four factors to be considered when selecting a standard setting method. The first consideration involves the format of the assessment. As stated previously, the format of the NAC OSCE exam does not permit the use of the test-centred methods generally used with educational assessments. The most appropriate holistic or compromise methods for use with the NAC OSCE exam have been outlined above in detail.

The second consideration for selecting a method is time. While there are not stringent constraints places on the time allotted to carry out the standard setting in the case of the NAC OSCE exam, those methods requiring an inordinate amount of time for training or carrying out such as the judgmental policy capturing method and the analytic method may not be appropriate. In the spring of 2013 the Hofstee method will be used with another exam, and because of this it may be of interest to compare this method to another for the purposes of this study. Lastly, when taking perceptions of the validity of the methods outlined into consideration, it appears the most commonly used methods in medical licensure currently are the borderline and contrasting groups methods. Because these appear most commonly in the literature, it may be advantageous to include one of the methods in the study for comparison.

The comparison of different methods has been studied, and results suggest that often different methods will produce different results (Cizek & Bunch, 2007). It is recommended that a method be chosen because it matches the format and purpose of the assessment. However, given the appearance of several methods that appear to match the needs for the NAC OSCE exam and the exploratory nature of the proposed study, a comparison of two or more methods seems both possible and appropriate.

Panelist Considerations

All of the methods discussed previously involve a panel of experts. The selection of a sufficiently large, representative panel is another major consideration of a standard setting study. In the case of the NAC OSCE exam, panelists should be practicing physicians who have experience supervising resident students, preferably International Medical Graduates. Hambleton et al. (2012) suggests a panel of 15-20 persons are "typical" in order to ensure the required diversity in terms of geographical location, age, gender, specialty, experience, etc. Jaeger (1991) suggests calculation of the required number of judges such that the standard error of the mean cut-score is small in comparison with the standard error of the test. Previous standard setting studies show the number of panelists ranging from as large as 17 (McKinley et al., 2005) to as small as 3-6 (Norcini et al., 1993).

In addition to stable estimates of cut-scores, the generalizability of the estimates is also a concern. That is, would another similar panel following the same methodology produce the same (or similar) results? In order to attempt to provide some evidence of generalizability of the cut-score, Hambleton et al. (2012) suggests selecting double the required panelists and having two parallel panels carry out the standard setting procedures. This design would allow for an investigation of the generalizability of the performance cut-scores set in the study. The use of two sub-panels is an excellent way to compile reliability data on any resulting cut score, and gives policy makers confidence that the resulting cut score has some generalizability across parallel forms and is not just a function of the idiosyncrasies of the panel.

Evaluating Validity

Kane (2001) provides three categories of validity evidence to be considered: procedural, internal and external. Procedural evidence has several sources upon which to draw: explicitness, practicability, implementation of procedures, panelist feedback and documentation (Hambleton & Pitoniak, 2006). While somewhat intuitive, the explicitness criterion requires that the steps of a standard setting should be made clear in order to allow for clear communication of the results of the study as well as to allow for replicability of the study (Hambleton & Pitoniak, 2006). Explicitness is evaluated by way of reviewing the extent to which the process and procedures to be carried out in the standard setting were defined prior to carrying out the study.

The practicability criterion allows for attention to be paid to issues of implementation and logistics (Berk, 1986). This criterion is evaluated by way of the ease of understanding of the method and results, and the intuitive nature of the method to the public (Hambleton & Pitoniak, 2006). In evaluating this criterion, another important consideration is the amount of time necessary in preparing for, training, and carrying out the standard setting.

Kane (2001) suggests evaluating the implementation of procedures criterion by way of the degree to which selection and training of panelists, definitions of performance levels, and data collection were performed systematically. Documentation regarding the required qualifications of panelists to participate as well as demographic information collected from the panelists themselves would serve as one piece of evidence of systematic implementation to show alignment between what was stated as required and the actual characteristics of the panel. Training and possible re-training in using the standard setting method will also allow for consistency throughout the meeting. In the case where there may be more than one panel participating in a single standard setting it may be advantageous to have all panelists participate in the same training to ensure a common understanding across the groups.

Kane (2001) also discusses the importance of panelist feedback and suggests evaluation of this criterion by way of systematic evaluation of the panelists' comfort and understanding of the process and cut-scores. This type of evidence can be collected easily and often throughout the standard setting. For example, surveys can be completed after orientation, training, rounds in the standard setting, and after completing the full standard setting meeting that will offer panelists to make note of any gaps in their knowledge before proceeding and to provide their feedback in terms of confidence in their ratings and preference of the method used.

Above all, the procedural evidence provides the information that is needed to properly document the process of the performance standard setting process from beginning to end. If cut-scores are questioned or challenged, the procedural evidence allows for a full account of measures undertaken ensuring fairness and validity of the method used and cut-scores set.

Internal sources of validity evidence include four sources concerned with consistency: consistency within method, intrapanelist consistency, interpanelist consistency, and other measures of consistency. Consistency with a method may be evaluated by two sources of information: the precision of the estimates of the cut-scores and across panel consistency (Kane, 2001). This estimate of consistency can be made via the use of two panels either consecutively or concurrently completing the standard setting using the same method, or using Generalizability theory to estimate the variance attributable to panelists and items.

Berk (1996) suggests evaluating intrapanelist consistency through two sources of evidence: the consistency of the panelists' ratings to empirical item difficulties and the consistency of individual panelists' ratings across rounds. However, there must be empirical data already collected and available for this use in order to assess alignment of ratings with item difficulties and so in some cases this is not possible. Additionally, it must be noted that there is going to be some change in ratings across rounds due to discussion and the presentation of feedback or impact data, but it is extreme change that will be considered anomalous.

Conversely, interpanelist consistency may be evaluated by way of consistency across panelists' ratings (Berk, 1996). Although it is expected that panelists will differ on their ratings of any single item or candidate, those ratings that are very different or inconsistent overall should be noted and discussed if possible. Other possible measures of consistency could be assessed across item type or content area depending on the individual assessment (Hambleton & Pitoniak, 2006).

Finally, external validity evidence of cut-scores concerns the consistency of cut-scores set with some external criteria. Though it is expected that the results of one method will differ from other possible methods, if there are drastic difference across methods or other external variables this can be evidence that the method used or procedures were biased in some way (Kane, 2001). Other comparisons that may be made are between the cut-score set and other sources of information such as GPA or academic references. Lastly, the extent to which resulting cut-scores are reasonable should be evaluated via impact or pass rates, and alignment with policy considerations (Kane, 2001).

As discussed, thorough documentation of the complete standard setting process is an important component of providing all types of validity evidence to be evaluated (Downing, Tekian, & Yudkowsky, 2006; Hambleton et al., 2012). A list of technical documentation as suggested by Pitoniak & Morgan (2012) includes: rationale for the standard setting method chosen, panelist recruitment and qualification information, study agenda, facilitator scripts, and an evaluation of the validity evidence.

Section 3. Practical Considerations

This section outlines the practical considerations in carrying out the proposed study. First, three different study designs are proposed each with possible positive and negative consequences provided. Second, a discussion of required training for the study is presented, including contingencies dependent on the design carried out. Third, following from the review of the literature, the three methods recommended for comparison are discussed including details regarding the minimal number of cases to be reviewed and the number of rounds. Finally, a list of required materials is included for the purposes of carrying out the study.

Possible Designs

Given the recommended comparison of three methods, there are three possible research designs that carry with them different advantages and disadvantages. The first design involves a single large (up to 15 panelists) completing all three methods as a group, with the same trainer and facilitator throughout. This design does not allow for any evidence of generalizability across panels. However, the advantage here is the all panelists will consistently see and experience the same training and facilitating throughout the standard setting. That is, there will be no confounding the experiences of the panelists. Also, a single panel would allow for a larger group and consequently, a more stable cut-score set. However, the design carries with it several limitations, notably, that methods two and three are not independent of the first method.

The second design involves two parallel panels that are sufficiently large (8-12 panelists each), with both panels completing the three methods in the same order. Training would be completed as a larger group with both panels being trained by the same person and breaking apart to complete the standard setting. However, the while the panels would have the same trainer, they would have different facilitators and we may see some effect due to that difference. The advantage of this design is that the multiple panels allow for evidence of generalizability across panels. However, the problem of confounded methods remains a problem as in the first design.

Finally, the third design involves two parallel panels that are sufficiently large (8-12 panelists each), controlling for order effects by switching the order of the presentation of the three methods. As with the second method, training would be completed are a larger group by a single trainer but the panels would have different facilitators for setting the actual cut-scores. This design helps to control for order effects and allows for generalizability evidence to be collected. However, we may see a facilitator effect as with the second design. Proper training of facilitators would mitigate this effect to some extent.

In general, several possible shortcomings must also be considered when designing the study (Boulet et al., 2003). Firstly, there can be the problem of case specificity in the sample of cases viewed by panelists. If the cases are not representative of what is seen on an exam they may not generalize to other forms. Second, biases can unintentionally inform the standard setting that are irrelevant to the actual examination. And finally, examinee-centred methods tend to require more time than test-centred methods. Viewing video performances or reviewing booklets or score profiles is much more time consuming than reviewing test items or rubrics and providing probabilities of correct responses. Of course, the actual time comparison would depend on the details of the study. For example, time can be controlled by limiting the number of videos that are used, or the number of available test booklets.

In this study all three of these concerns will be addressed in the design and execution of the standard setting in order to mitigate their effects. The cases and MCQ items chosen for use in the standard setting will be an entire form of the exam allowing for diversity in both content and complexity. Equating across forms and administrations of the NAC OSCE exam will also allow for the cut-score to be more generalizable. The individual candidate profiles used for the study will also be representative in terms of race, gender, language proficiency, and age.

Finally, though the use of video recording of candidate performances may be preferable to panelists, the time it would take in order to review an adequate number of performances is prohibitive. Therefore, videos will be available to panelists in the event that they need more information about a candidate to make a decision. Also, panelists will be able to fast forward through videos if they only need to view a small portion of the performance to make their decision. This will allow panelists to move through candidate profiles more quickly and efficiently.

Training

To begin training, all panelists must be familiarized with the format and purpose of the NAC OSCE examination. This includes a description of the types of candidates seen for the exam as well as the consequences of a pass/fail decision for the candidates. All panelists should complete the therapeutics component of the exam as well as at least one OSCE station. Panelists should be expected to complete the OSCE case and serve as an examiner for another panelist. These are common practices in standard setting studies.

Next, a discussion surrounding the performance level descriptions as provided by the Medical Council of Canada should be complete and detailed. This includes consensus on the description of the performance categories that decisions will be based upon. This step will include the viewing of performance samples (including completed rating sheets and guidelines to rating descriptors (Appendix A), scored MCQ test booklets (Appendix B), and case information for examiners (Appendix C) which includes specific expectations for competency mastery) and discussion of the score profiles of 5 (ranging from clear unacceptable to clear acceptable performance) candidates. Note: this paper provides placeholders for documents that would be included in the standard setting materials provided for panelists. This allows the group to classify and discusses aberrant classifications to obtain full and complete descriptions of the MCC.

If parallel panels are to be used, it would be preferable that panels are trained together and then separated as we know the importance of consistent and comprehensive training to be an integral component in a standard setting (Boulet et al., 2003; De Champlain, 2004; McKinley et al., 2005). Extensive training of panelists before moving forward with the standard setting will help them to focus on the true purpose of the exam instead of extraneous details that may unduly influence their ratings. Training provides an opportunity for panelists to ask questions and have initial discussions allowing orientation to the task at hand.

Methods Recommended

Contrasting Groups Method

Performance sample information is provided to panelists including completed rating sheets, scored MCQ test booklets, guidelines to rating descriptors and case information for examiners which includes specific expectations for competency mastery. Panels A and B will review the same candidate profiles and MCQs. For the therapeutics component, panelists will be asked to independently classify at least 50 candidates as passing or failing. Once classifications have been made, those candidates where panelists do not agree can be discussed and ratings may be changed. At the end of round 1, the panel will be provided with a depiction of the two score distributions (pass/fail) and a cut-score to be determined via logistic regression. A second round will allow for judges to discuss and reconsider classifications if they so choose.

Logistic regression is a type of analysis that is used for the prediction of a dichotomous outcome (in this case, pass/fail) based on one or more predictor variables. The analysis identifies a raw test score at which the likelihood of being assigned to a category (pass, in this case) is equal to or exceeds the likelihood of being assigned to a lower level (fail). Figure 1 shows the hypothetical score distributions showing a potential cut-score. We can see in the figure that a candidate with a raw score of 5 is equally likely to be classified as passing as he/she is of failing.

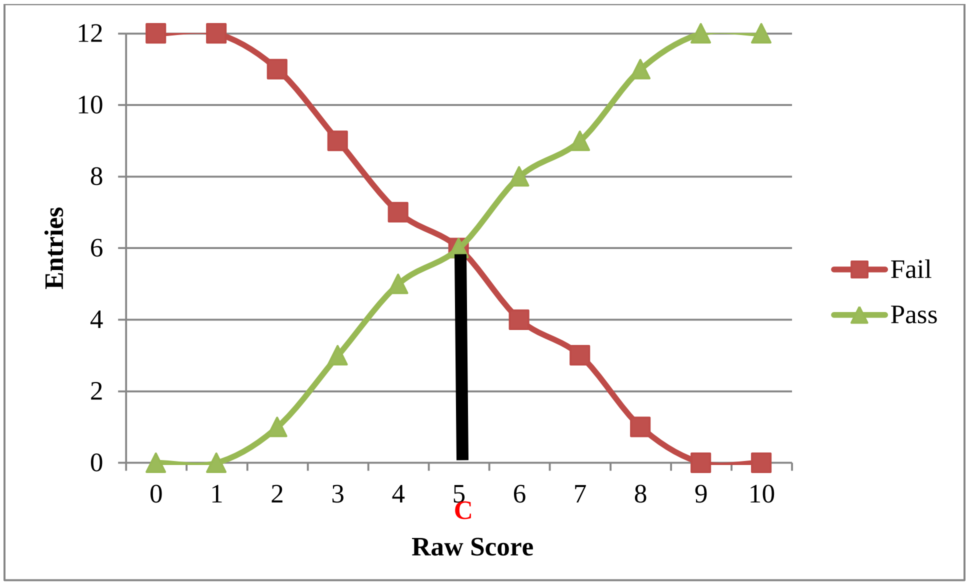

Figure 1. Graphical display of hypothetical score distributions showing a potential cut score at the intersection of the adjacent distributions.

For the OSCE component of the examination, a minimum of 30 candidates will be reviewed on 10 cases; 10 of which are clearly pass, 10 are clearly fail, and 10 are borderline. In the interest of time, videos of all performances will be made available to panelists to review but the review is not mandatory. It will be recommended that video review occur where consensus cannot be reached without further information. Panelists will be allowed to fast-forward videos or watch only a portion if that is what is necessary for a classification decision. However, rewinding or re-watching portions of the performance will not be permitted because this is inconsistent with the experience of Physician Examiners during the actual exam.

As in the case of the therapeutics component, panelists will independently classify candidates as either pass or fail based upon their profiles and those candidates that panelists do not agree on may be discussed. After the first round the panelists will be provided depictions of score distributions based upon their decisions as well as a cut-score determined by logistic regression. The groups will be then given an opportunity to revisit their classifications.

Body of Work Method

Work samples including the same kinds of information as in the contrasting groups method are ordered by total score (low to high) and provided to panelists. A minimum of 30 profiles with the same OSCE cases and MCQs should be reviewed, with scores across the score scale. Panels A and B will have the same available samples of work, but these samples will differ from the samples used in the contrasting groups method. Similarly, videos of all performances will be made available to panelists to review but the review is not mandatory. It will be recommended that video review occur where consensus cannot be reached without further information. Panelists will be allowed to fast-forward videos or watch only a portion if that is what is necessary for a classification decision. However, rewinding or re-watching portions of the performance will not be permitted because this is inconsistent with the experience of Physician Examiners during the actual exam.

Typically, this method includes two rounds. In the first round, called rangefinding, panelists examine the work samples and classify as pass/fail. Those samples wherein the entire panel agreed upon pass/fail classification are eliminated from the samples and discussion allows for panelists to change their ratings. The second round, called the pinpointing round, begins with work samples with total scores close to those remaining from the rangefinding round being added. Profiles are again classified as pass/fail and those agreed upon by the panel are eliminated. Remaining samples are discussed and panelists are given an opportunity to change their ratings. Logistic regression is used in the calculation of the final cut-score.

Hofstee Method

As the method that will be used for another OSCE examination program at the Medical Council of Canada, and the method is relatively efficient, it will also be compared to the other two methods used. Immediately following the completion of the body of work method, panelists will be asked four questions:

1. What is the highest percent correct cut-score that would be acceptable, even if every examinee receives that score?
2. What is the lowest percent correct cut-score that would be acceptable, even if no examinee receives that score?
3. What is the maximum acceptable failure rate?
4. What is the minimum acceptable failure rate?

The mean values provided by panelists for the four questions will be calculated and points will be plotted on a graph with the axes “percentage failing” and “percentage correct required” based on a cumulative distribution of total scores. Point one corresponds to the mean values given for questions 1 and 4 and point two corresponds to the mean values given in response to questions 2 and 3. Along the straight line between points 1 and 2 are all possible values of cut-scores. An observed test distribution is projected onto the same graph and where the straight line between points 1 and 2 intersects with the distribution becomes the cut-score. Because real data is available, pass/fail rates can be provided to the panels and they will be given an opportunity to modify their responses. Figure 2 illustrates a Hofstee plot with a hypothetical test score distribution. The four coordinates for the two points on the plot coincide with the four questions asked of the panel. Along the line connecting the two points are all “compromise solutions”, and the optimal solution is the point at which the this line intersects with the test score distribution indicated by the “X” on the plot.

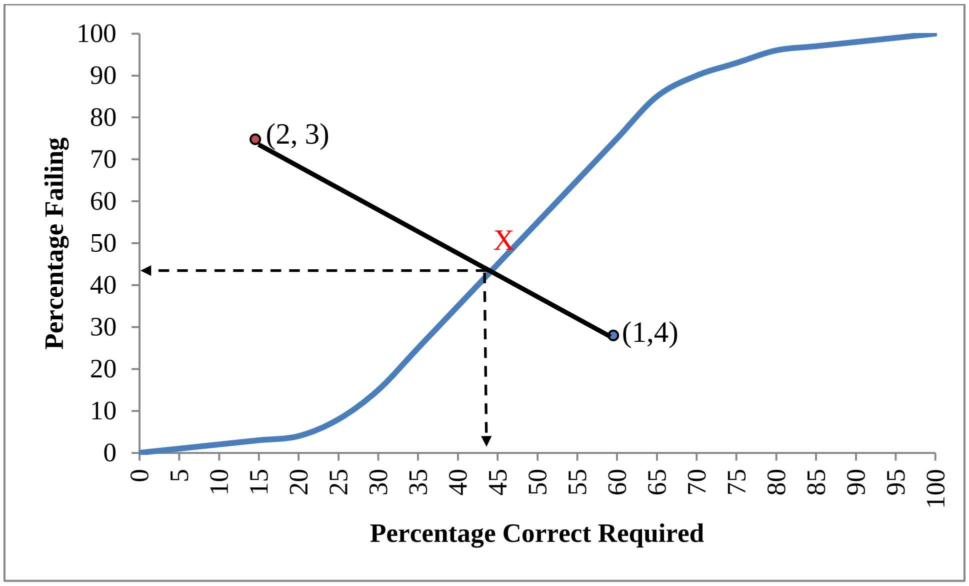

Figure 2. Hypothetical illustration of a Hofstee cut-score derivation.

Materials

This section provides a list of the materials recommended for carrying out the standard setting study comparing three methods. Some of the materials needed will be independent of the methods used and others will be required specifically for each method; the list is divided to highlight which are independent and which are specific. Examples of several of the materials are provided in the Appendices.

Required materials independent of the method used:

- A meeting agenda.
- Performance level descriptors.
- A full description of the “just qualified” candidate.
- Performance Samples including the information contained in Appendix A, B, and C.

- If videos are to be used, videos must be available matching the performance samples provided.
- Orientation Survey (Appendix D; Cizek, 2012b).
- Within-Method Evaluation Surveys (Appendix F, G; Cizek, 2012b).
- Pre-Demographic Survey including educational and employment background, experience with the NAC OSCE examination, and IMGs.

Required materials specific to the method:

- Post-Training Surveys for each method (Appendix E; Cizek, 2012b).
- Post-Method Survey for each method (Appendix H; Cizek, 2012b).
- Rating forms required for each method.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508–600). Washington, DC: American Council on Education.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, *15*, 4–9.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*, 137–172.

Berk, R. A. (1996). Standard setting: The next generation (where few pschometricians have gone before!). *Applied Measurement in Education*, *9*, 215–235.

Boulet, J. R., De Champlain, A. F., & McKinley, D. W. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical teacher*, *25*(3), 245–9. doi:10.1080/0142159031000100274

Boursicot, K. a M., Roberts, T. E., & Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical education*, *41*(11), 1024–31. doi:10.1111/j.1365-2923.2007.02857.x

Cizek, G. J. (2012a). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (Second Edi., pp. 3–14). New York, NY: Routledge.

Cizek, G. J. (2012b). The forms and functions of evaluations in the standard setting process. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (Second Edi., pp. 165–178). New York, NY: Routledge.

Cizek, G. J. (Ed.). (2012c). *Setting performance standards: Foundations, methods, and innovations* (Second Edi., p. 588). New York, NY: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests* (p. 352). Thousand Oaks, CA: Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards : Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50. doi:10.1111/j.1745-3992.2004.tb00166.x

Dauphinee, W. D., Blackmore, D. E., Smee, S., Rothman, A. I., & Reznick, R. (1997). Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Advances in Health Sciences Education*, *2*(3), 201–211. doi:10.1023/A:1009768127620

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*(1), 62–66. Retrieved from http://www.utpjournals.com.login.ezproxy.library.ualberta.ca/jvme/tocs/311/61.pdf

Downing, S. M., Tekian, A., & Yudkowsky, R. (2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and learning in medicine*, *18*(1), 50–7. doi:10.1207/s15328015tlm1801_11

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting Performance Standards on Complex Educational Assessments. *Applied Psychological Measurement*, *24*(4), 355–366. doi:10.1177/01466210022031804

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth Edi., pp. 433–470). Westport, CT: Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing reliability of results. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (Second Edi., pp. 47–76). New York, NY: Routledge.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Appled Measurement in Education*, *8*, 45–55.

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

Humphrey-Murto, S., & MacFayden, J. C. (2002). Standard setting: A comparison of case-author and modified borderline group methods in a small-scale OSCE. *Academic Medicine*, *77*(7), 729–732.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (Third Edit., pp. 485–514). New York, NY: Macmillan.

Jaeger, Richard M. (1991). Selection of Judges for Standard-Setting. *Educational Measurement: Issues and Practice*, *10*(2), 3–14. doi:10.1111/j.1745-3992.1991.tb00185.x

Jaeger, R. M. (1995a). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice*, *14*(4), 16–20.

Jaeger, R. M. (1995b). Setting performance standards through two-stage judgemental policy capturing. *Applied Measurement in Education*, *8*, 15–40.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 53–88). Mahwah, NJ: Erlbaum.

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, I. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods and perspectives* (pp. 219–248). Mahwah, NJ: Erlbaum.

Kingston, N. M., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In Gregory J Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (Second Edi., pp. 201–223). New York, NY: Routledge.

Kramer, A., Muijtjens, A., Jansen, K., Dusman, H., Tan, L., & Van Der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, *37*, 132–139. doi:10.1046/j.1365-2923.2003.01429.x

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In Gregory J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (Second Edi., pp. 225–253). New York, NY: Routledge.

Livingstone, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on Educational and occupational tests*. Princeton, NJ.

McKinley, D. W., Boulet, J. R., & Hambleton, R. K. (2005). A work-centered approach for setting passing scores on performance-based assessments. *Evaluation & the health professions*, *28*(3), 349–69. doi:10.1177/0163278705278282

Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 219–252). Lincloln, NE: Buros Institude of Mental Measurements.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3–19.

Norcini, J. J., Stillman, P. L., Sutnick, A. I., Regan, M. B., Haley, H. L., Williams, R. G., & Friedman, M. (1993). Scoring and standard setting with standardized patients. *Evaluation & the health professions*, *16*(3), 322–332.

Smee, S. M., & Blackmore, D. E. (2008). Setting standards for an objective structured clinical examination: The borderline group method gains ground on Angoff. *Medical Education*, *35*(11), 1009–1010. doi:10.1111/j.1365-2923.2001.01047.x

Appendices

Appendix A

Rating Scale Form

Appendix B

Sample Therapeutics Examination

Appendix C

Case Information for Examiners

Appendix D

Evaluation #1: Evaluation of Orientation Session

Directions: Please check one box for each of the following statements by placing an “X” in the box corresponding to your opinion.

SD- Strongly Disagree; D- Disagree; N- Neutral; A- Agree; SA- Strongly Agree

	Statement	**SD**	**D**	**N**	**A**	**SA**
1	The orientation session provided me with a clear overview of the purpose of the standard setting for the National Assessment Collaboration (NAC) Exam.					
2	The orientation session answered questions I had about standard setting for the NAC OSCE exam.					
3	I have a good understanding of my role in the standard setting activity.					
4	Reviewing the NAC OSCE exam content helped me to understand the standard setting task.					
5	Experiencing the NAC OSCE exam helped me understand the difficulty, content, and other aspects of the examination.					
6	I have a good understanding of the performance level descriptors (PLDs).					
7	Examining actual samples of examinee work was helpful.					
8	I felt comfortable contributing the in the group discussions.					
9	The facilitators helped me understand the standard setting process.					
10	The meeting facilities and materials have been conducive to accomplishing the tasks.					
11	The timing and pace of the orientation were appropriate.					

12 One thing that might require additional explanation before we move on is:_______________

____

13 Other comments or suggestions:____

____

Thank you for completing this survey.

Appendix E

Evaluation #2 – End of Method Training

Directions: Please check one box for each of the following statements by placing an “X” in the box corresponding to your opinion.

SD- Strongly Disagree; D- Disagree; N- Neutral; A- Agree; SA- Strongly Agree

	Statement	**SD**	**D**	**N**	**A**	**SA**
1	I have a good understanding of the borderlines between the performance levels.					
2	The training in the standard setting method was clear.					
3	The practice using the standard setting method helped me to understand how to apply the method.					
4	I am comfortable with my ability to apply the standard setting method.					
5	I understand the kinds of feedback that will be provided to me during the standard setting process.					
6	The timing and pace of the method training were appropriate.					
7	Overall, I feel prepared to begin the standard setting task.					

8 One thing that might require additional explanation before we move on is:_______________

____

9 Other comments or suggestions:____

____

Thank you for completing this survey.

Appendix F

Evaluation #3 – Round 1

Directions: Please check one box for each of the following statements by placing an “X” in the box corresponding to your opinion.

SD- Strongly Disagree; D- Disagree; N- Neutral; A- Agree; SA- Strongly Agree

	Statement	**SD**	**D**	**N**	**A**	**SA**
1	I understood how to complete my round 1 ratings.					
2	I am confident in my round 1 ratings.					
3	I had the opportunity to ask questions while working on my round 1 ratings.					
4	The facilitators helped to answer my questions while working on round 1 ratings.					
5	The technologies were helpful and functioned well.					
6	The timing and pace of round 1 activities was appropriate.					

7 One thing that might require additional explanation before we move on is:______________

____

8 Other comments or suggestions:____

____

Thank you for completing this survey.

Appendix G

Evaluation #4 – End of Round 2

Directions: Please check one box for each of the following statements by placing an "X" in the box corresponding to your opinion.

SD- Strongly Disagree; D- Disagree; N- Neutral; A- Agree; SA- Strongly Agree

	Statement	**SD**	**D**	**N**	**A**	**SA**
1	The discussion of the round 1 ratings and instructions helped me to understand what I needed to do to complete round 2.					
2	I understood how to complete my overall cut-score recommendations.					
3	I am confident in my overall cut-score recommendations.					
4	I had the opportunity to ask questions while working on my final recommendations.					
5	The facilitators helped to answer questions and to ensure that everyone's input was respected and valued as we worked on our final recommendations.					
6	The timing and pace of the final round was appropriate.					

7 One thing that might require additional explanation before we move on is:______________

____

8 Other comments or suggestions:____

____

Thank you for completing this survey.

Appendix H

Evaluation #5 – Post-Method Evaluation

Directions: Please check one box for each of the following statements by placing an "X" in the box corresponding to your opinion.

SD- Strongly Disagree; D- Disagree; N- Neutral; A- Agree; SA- Strongly Agree

	Statement	**SD**	**D**	**N**	**A**	**SA**
1	Overall, the facilities and food service helped to create a good working environment.					
2	The technologies were helpful and functioned well.					
3	Overall, the training in the standard setting purpose and methods was clear.					
4	Overall, I am confident that I was able to apply the standard setting methods appropriately.					
5	Overall, the standard setting procedures allowed me to use my experience and expertise to recommend cut-scores for the NAC OSCE exam.					
6	Overall, the facilitators helped to ensure that everyone was able to contribute to the group discussions and that no one unfairly dominated the discussions.					
7	Overall, I was able to understand and use the information provided.					
8	The final group-recommended cut-score for the pass level fairly represents the minimal level of performance for examinees at the pass level.					
9	If you answered D or SD to Q8, do you believe the final group-recommended cut-score is: __ too high or __ too low? (check one)					

10 One thing that might require additional explanation before we move on is:_______________

____

11 Other comments or suggestions:____

____

The list below contains the sources of information available for generating your ratings during the standard setting process. Please place an "X" in the box following the source to indicate how much you relied on that source of information in making your judgement. Please mark only one "X" in each row.

Next, consider which source of information you relied upon most, and which you relied upon least, to make your judgements. Place one "+" in the column at the far right to indicate the one source of information you relied upon most, and one "-" to indicate the source you relied upon least.

	Statement	**Heavily**	**Moderately**	**Slightly**	**Not at All**	**+/-**
12	My experience taking the test.					
13	My own experiences with students.					
14	The performance level descriptors (PLDs).					
15	The descriptions of the borderline groups.					
16	The samples of candidate performance.					
17	The videos of candidate performances.					
18	The group discussions.					

19 Other comments:____

____

Thank you for completing this survey.