Medical Council of Canada
Research and Development Report

Evidence for the Validity of a Clinical Skill Assessment

Timothy J. Wood, Marguerite Roy, Meghan M. McConnell, Krista Breithaupt

DRAFT

When developing and administering a high stakes licensing examination, program administrators and our candidates require us to seek and provide evidence that the decisions made using test scores are justified. Often this evidence is discussed in the context of the validity of test score use. The notion of validity in this context can be understood as evidence supporting the use of scores for their intended purpose (Messick, 1989; AERA, NCME & APA, 1999, Kane, 2006).

Medical Council of Canada (MCC) is developing a clinical skills examination for use with international medical graduates (IMGs) wishing to enter residency training in Canada. The program is being created in accordance with best practices for national licensing exams to guide design, scoring, and reporting to help ensure legal defensibility and high quality. The purpose of this report is to use data from a pilot of this new clinical skills examination to illustrate a validity framework that reflects contemporary measurement theory and is practical for use in exams for medical licensure.

In Canada and the United States, all medical schools undergo an accreditation process to ensure that graduates have a consistent level of training and knowledge in medicine. International medical schools do not undergo the equivalent process. This means the comparability of international programs is difficult to assess against the North American standard. To evaluate the skills and knowledge of IMGs, seven of the provinces in Canada have set up IMG assessment programs. The goal of these programs is to determine what is an appropriate training or scope definition as they proceed towards licensure. Each provincial assessment program has been created to meet the needs of their respective province but this has resulted in a set of assessments that vary across provinces. Even if the same examination format is used, the content, scoring, and overall design is not standardized across examinations. This variety across assessment programs results in provinces not recognizing the results of the assessment programs from other provinces and therefore has an impact on the IMG who is seeking licensure. For example, an IMG seeking licensure in both Ontario and Quebec would need to complete two assessments programs to be considered for licensure both of which can be quite different and quite expensive to take.

The National Assessment Collaboration (NAC) was formed as alliance of provincial IMG programs and related stakeholders seeking to streamline and standardize the evaluation process used to assess IMGs who wish to obtain a license to practice medicine in Canada. As a first step in the assessment process, MCC was asked by the NAC to create a national objective structured clinical examination (OSCE) that could be used to assist clinical program directors with the selection of IMG applicants into their individual postgraduate training programs. The intent of this examination was to

create a centrally managed program to ensure standardized processes, while enabling locally convenient administration across Canada. The examination development began in 2009, piloting began in 2010 and the program is intended to become fully operational in the next two years.

An OSCE for entry to postgraduate training was chosen as a first step in the NAC assessment format for a reason. When selecting IMG residents for post graduate training in Canadian medical schools, applicants, program directors will consider the applicants' qualifications including which medical school they graduated from, their performance, and their interests. In North America, because the medical schools undergo an accreditation process, the program directors know that graduates of Canadian or US medical schools have a consistent level of training and knowledge in medicine. International medical schools do not undergo a similar accreditation process so when selecting an IMG into a training program, the program director always has some uncertainty as to the level of training the IMG has received especially when compared to North American graduates. By implementing a standardized national OSCE to assess the readiness of the IMG to enter postgraduate training, it is hoped that program directors will use this information when selecting IMGs to their program.

An OSCE is an examination in which examinees rotate sequentially through a series of "stations" at which specific clinical tasks have to be performed. The length of the stations is usually short (e.g., five to ten minutes) and the tasks at these stations could involve taking a history from a patient, doing a physical examination of a patient, managing a problem, or performing a technical skill. In most cases, the station will have an actor (standardized patient) playing the role of a patient and the actions within the station are scripted so that the patient role is played as consistently across examinees and across performances as possible. Some OSCEs use the standardized patient to score the examinee and some OSCEs use an examiner (often a physician). The marking scheme for each station is structured in advance and will consist either of a checklist, a rating scale or a combination of both.

The OSCE has arguably become the most common method used in medicine to assess clinical skills and considerable research has been published on the measurement properties of the examination (for reviews see the following: van der Vleuten & Swanson, 1990; Swanson, Clauser, and Case 1999; Petrusa 2002; Boulet, Smee, Dillon and Gimpel, 2009). While validity issues have certainly been at the forefront of much of this research, the focus has usually been on testing different types of validity (e.g. content validity, construct validity, predictive validity etc). It is only recently that OSCE research, and in fact much of medical education research has started to talk in terms of a unified theory of validity

(Downing and Haladyna, 2009; Beckman and Cook 2005; Clauser, Margolis, and Swanson, 2008; Margolis, Clauser, Winward, and Dillon, 2010; Zumbo et al, xxx).

Some frameworks for validity currently discuss how test scores are interpreted; that is, what are the inferences one can make about a candidate based on the score from the examination (e.g. Kane 2006; Messick, 1989; AERA et al., 1999; Downing & Haladyna, 2009; Hubley and Zumbo, in press). The process of validation research usually involves collecting multiple sources of evidence that supports the proposed interpretation of a test result. The evidence can be grouped into different sources (Clauser et al., 2008) and the lexicon of various theorists may differ somewhat arbitrarily. Information from a variety of sources is usually used to support judgment about the meaning of the test score. Table 1 displays five sources of validity evidence to be used in this study of an OSCE, based on the 1999 Standards (AERA, NCME, APA, 1999).

Developing a new examination is a multi-step process with many critical decisions. During the pilot phase of this program, MCC administered three examinations (two forms) that were structured similarly to the MCC Qualifying Examination Part II (MCCQE Part II) clinical skills model. The MCCQE Part II is high-stakes OSCE used as a requirement for medical licensure in Canada. The difficulty of the MCCQE Part II is established at the level of someone who is ready for unsupervised practice; that is, a person who has completed one year of postgraduate training so is at a more difficult level than is intended for the NAC, however, adjustments were made to the standard setting to take this difference into account. Lessons learned from these administrations informed the development of the new NAC OSCE which will be implemented in March 2011.

For the current manuscript, the evaluation of validity evidence for the NAC program will be discussed, as will the results of this proof of concept examination and implications for the stabilized version of the program. This illustration will have some generalizability for other clinical skills or task evaluations used for assessment.

Table 1. *Five Sources of Validity Evidence*

Evidence	Description	Types of information
Content Evidence	Examines the relationship between content and what the assessment tool is supposed to be measuring	How well does a test blueprint match the purpose of the test and what is the match between stations and the blueprint; information on the characteristics of item-writers and how the stations were developed.
Response Process Evidence	Cognitive models that describe the relationship between the provided responses and the underlying analysis required to solve the challenge represented in the task. A cognitive theory for the task and response are normally posited as a basis for this analysis.	Features of the responses are often evaluated for support of the theory describing cognitive processes that are the basis for the solution. Timing data, work-aloud and categorization of tasks long a cognitive complexity scale are often useful support data.
Internal Structure Evidence	Examines the degree the items are related to what the assessment tool is supposed to be measuring, i.e., what are the statistical properties of the instrument	item analyses to identify poor performing stations; reliability to see if a score or pass mark is reproducible; generalizability analysis to examine sources of errors.
Relations to other variables	Examines the relationship between scores on this assessment tool and other measures that are supposed to be related or unrelated	Correlations with other similar measures and dissimilar measures to determine if the correlations are in the manner that would be expected.
Consequences	Examines the intended and unintended consequences of how the scores are used.	what is the impact of the scores and pass/fail decisions on stakeholders,

Methods

Participants

A total of 162 participants from three provincial IMG assessment programs attempted the two administrations of the NAC OSCE in 2010. The programs involved included the Collège des Médecin du Québec (CMQ), IMG-British Columbia (IMG-BC), and the Manitoba Licensing Program for IMGs (MLPIMGs)).

Design of the Administration

Two forms of the examination were created, each made up of 16 stations, 12 of which include a patient encounter and 4 of which contain a series of short answer questions that require written responses from candidates. At each of the 12 patient encounter stations, candidates interacted with a standardized patient (SP) and a physician examiner. The tasks performed by the candidate required them to take the medical history, conduct a physical examination, and/or diagnose and manage and complaints. Of 12 stations, six were ten-minute stations and six were longer because they included a written component after the case encounter. This written task is called a post-encounter probe (PEP), where candidates may be asked to document findings from the preceding patient encounter, provide differential diagnoses, write admission orders, or interpret laboratory results, x-rays, etc.

The remaining four stations were item requiring therapeutics knowledge. This component was deemed essential for the NAC OSCE, the provincial assessment programs noted IMGs typically show poor knowledge compared to Canadian- trained physicians and this can impede competent practice even in a supervised setting such as a residency program. At each therapeutics station, candidates were provided with a small booklet that contained three to seven therapeutics questions for a total of 24 scored items.

Blueprint Criteria and Content

For the 2010 examinations, two forms were created with 10 of the 12 stations being donated to the examination by MCC thereby ensuring that the quality of these stations matched that used on the licensing examination. The remaining two stations were donated by a partner assessment program and were edited by the test committee before being included on the examination. The blueprint for this pilot examination was based on the one used for the MCCQE, which is displayed in Table 2, and the stations were chosen to be match the blueprint for the MCCQE Part II as much as possible. As shown in this table, the stations included clinical problems relating to internal medicine, obstetrics and gynecology (OBGYN), surgery, pediatrics, and psychiatry. Other skills that were considered included history-taking, physical examination and counseling patients. Also SP age and SP gender are considered. The stations used on both forms of the examinations matched this blueprint except for gender. A total of 35% of the cases involved a male SP and 65% involved a female SP but this variance was deemed to be acceptable to the test committee.

Table 2. *Examination blueprint for the MCCQE Part II*

Discipline		Clinical Domain	
N of Stations	Description	N of Stations	Description
2-4	Medicine	2	Counseling
1-3	OBGYN	4-5	History
1-3	Pediatrics	2	management
1-3	Psychiatry	4-3	Physical
2-4	Surgery	all stations	patient Interaction (e.g. communication)

Body Systems		Age	
N of Stations	Description	N of Stations	Description
at least 1	gastrointestinal	at least 1	child 1 mo-12 yrs
at least 1	caridac	at least 1	Adolescent 13-18 yrs
at least 1	musculoskeletal	at least 1	Aged 19-44 yrs
at least 1	pulmonary	at least 1	Aged 45-64 yrs
		at least 1	Aged 65+

Other Categories	
SP Gender	Ratio of male to female cases should be no greater than 40/60 in either direction

In addition to patient encounter stations, the NAC OSCE also included a written component designed to assess candidates' understanding of basic therapeutic knowledge. The questions were donated to MCC for use on the NAC OSCE by the Clinical Assessment for Practice Program (CAPP) in Nova Scotia, which uses these questions on a therapeutics examination that they administer to IMGs who believe they are ready to practice medicine in Nova Scotia. The blueprint categories used for the therapeutic component were derived from the blueprint of the CAPP examination with the numbers adjusted to match the total of 24 questions that were used. This blueprint is displayed in Table 3.

Table 3. *Examination blueprint for the Therapeutics Component of the 2010 NAC OSCE*

Therapeutics Blueprint						
	Age Group					
Therapeutic option	Infants / children	Adolescents (13-17)	Adults (18-60)	Adults (>60)	None	TOTAL
Pharmacotherapy	2	2	7	2	1	14 (58%)
Adverse Effects	1		2	1		4 (17%)
Disease Prevention	1		2	1		4 (17%)
Health Promotion			2			2 (8%)
TOTAL	4 (17%)	2 (8%)	13 (54%)	4 (17%)	1 (4%)	24 (100%)

Scoring

Scoring for this examination was identical to that used on the MCCQE Part II. During clinical encounters, physician examiners were asked to observe the candidate as they interacted with the patient and to complete a checklist of actions. These checklists consist of sets of tasks that candidates are expected to perform while interacting with a patient. Checklist items are weighted in terms of their importance and a station score is created by converting the sum of the checklist scores to a percentage. In addition to the checklists, examiners are asked to complete a holistic six-point rating of overall quality as expected by someone entering their first year of residency. The rating scale for the scale was defined as; excellent, good, borderline satisfactory, borderline unsatisfactory, poor, or inferior. The holistic rating scale was only used for establishing a pass or fail on the total score attained by examinees using a procedure called the modified borderline method. For each station, those examinees who are identified as having a borderline performance are identified (either borderline fail or borderline pass) and the mean checklist score for these candidates is defined as the pass mark for the clinical portion of the station. For the PEPs and therapeutic questions, a team of graders scored the questions according to the answer key and graders were asked to complete a holistic rating similar to the clinical encounter. Scores on the PEPs were converted to percentages and were combined with the checklist items (50% weighting for both) from the clinical encounter to produce a total station score for the couplet stations. Scores on the therapeutic station were also converted to percentage scores and combined to produce a therapeutics score for the examination. The final pass/fail mark for the examination was the sum of the 12 station cut scores (weighted 75%) and the sum of the therapeutic cut scores (weighted at 25%).

Results

Content Evidence

Content evidence requires an appraisal of expert categorization of tasks and scored items against the construct as defined by the approved blueprint of knowledge and skills to be assessed. In this study, the approved forms included cases that were carefully reviewed and revised by the policy group of experts appointed as the program test committee. The cases were chosen to match the blueprint, and were also identical to those used in the MCCQE Part II with standard setting revised for entry into residency.

Response Process Evidence

This type of validity evidence deals with the responses from the examinees and how their performances are captured in the scoring of each task. A link to the cognitive activities needed to solve sophisticated problems is usually based on a theoretical framework. Sometimes supporting evidence can come from surveying examinees to determine strategies they use to guide performance, or surveying raters to determine how they assign a mark. Studying how relevant subgroups differ in their responses would also be considered to be part of this type of evidence.

Because of the similarity between the 2010 NAC examinations and the MCCQE Part II, many of the testing procedures from the latter examination were implemented. For example, all SP and examiner training procedures were borrowed from the MCCQE Part II which would ensure a standardized training protocol was implemented across and within sites. Also, all data quality assurance procedures used on the MCCQE Part II were implemented which ensured the integrity of the data. This pilot examination did not delve into a further exploration of the rationale behind the scoring procedures but this issue was flagged for further study.

Internal Structure

An analysis of the internal structure of the response data can be used as one source of evidence that scores on the test items and relationships between those responses are consistent with what would be predicted based on the constructs being measured. Information that is collected could range from estimators that measure the correspondence between demonstrations of related skills and knowledge, and evidence for discrete demonstrations of an underlying trait or ability.

Item analyses were conducted on station scores, only one couplet station was flagged due to a low item-total correlation (item-total correlation below 0.20) indicating poor representation of a common construct. Upon review of response patterns for the same content in a case from the MCCQE Part II, it was discovered that proportion correct was high. This finding could be expected to be paralleled so that little variance in the case score would result in a low correlation to the total scores obtained by a candidate. No other stations were flagged for poor performance; all stations had item-total correlations above 0.20. Table 4 displays the descriptive statistics and acceptably high reliability estimates for total scores on each administration (Cronbach's alpha). Mean total scores were also similar on the forms, and there was a slight difference in the mean scores for the therapeutic examination. It may be important to note Form 1 was not used for pass/fail purposes by the local program and a cut score was not estimated.

Table 4. *Descriptive statistics (in percentages) for the NAC OSCE and for two recent administrations of the MCCQE Part II.*

	n	Total Score	SD	Therapeutic	Cut score	Pass Rate	Alpha
NAC Form 1	70	62	7.3	57	n/a	n/a	0.77
NAC Form 2	92	62	7.5	66	66	78	0.80
MCCQE Part II fall	2644	70	6.1	n/a	58	78	0.73
MCCQE Part II spring	977	62	7.1	n/a	58	65	0.75

Relations to Other Variables

An analysis of the relationship between test scores to other variables outside of the test that may be related or unrelated to the construct being measured is an important step to support the hypothesis that results reflect a distinct knowledge or skill domain. Table 4 also displays the descriptive statistics for two recent administrations of the MCCQE Part II. As shown in the table, the mean total score for the NAC OSCE administrations tended to be similar or slightly below mean score for the MCCQE Part II where as the cut scores tend to be higher. This pattern would be expected given that the

MCCQE Part II content, passing score, and scoring guidelines are designed to assess readiness for unsupervised practice where as the NAC OSCE is designed to assess readiness for residency.

Table 5 displays a summary of the demographic characteristics of candidates who took the NAC OSCE and those who took two recent administrations of the MCCQE Part II. The purpose of this table is to determine how similar the cohorts were for both examinations which will help interpret the pattern of scores found in Table 4. Note that this information was not available for candidates who attempted form 1. As shown in the table, the cohort that took the NAC OSCE tended to be more similar to the cohort that attempted the spring MCQE Part II than the fall examination. It is also interesting that in Table 4, the mean score for the spring examination is similar to the score for the NAC OSCE whereas the mean score for the fall MCCQE Part II is higher. Given that the fall examination consists of a high proportion of Canadian medical graduates (approximately 70%) whereas the spring examination consists of a higher proportion IMGs (approximately 70%), the pattern found in Table 4 supports the idea that mean performance is a result of the skill and ability of this IMG population.

Table 5. *Demographic characteristics of candidates who attempted the NAC OSCE and two recent administrations of the MCCQE Part II.*

	Mean Age	Mean years since graduation	% female
MCCQE Part II fall	32	5	52
MCCQE Part II spring	37	10	44
NAC Form2	40	7	36

Consequences

Intended and unintended consequences of a test were considered as validity evidence associated with the pilot test scores. Information in this area could include the impact of a pass/fail decision on candidates, or how assessment programs make use the results of the examination. Authors will explore the available evidence that residency directors used rankings and pass or fail status to admit candidates. This is only an initial source, it would be reasonable to follow up with post-graduate program directors and regulators to explore implications of the standings.

Discussion

When developing and administering any examination, it is important to examine a variety of sources of support that scores from the assessment are interpreted as intended. While issues related to validity have been discussed in medical education, it seems contemporary views of validity are just now being considered (Downing and Haladyna, 2009; Beckman and Cook 2005; Clauser et al., 2008; Margolis et al., 2010; Zumbo et al, xxx). The purpose of this report was to document a procedure that could be used to explore the validity of a new high-stakes clinical examination. As such, the types of information that can be collected during the development of a new OSCE were reported within the context of a unified model of validity based on the current Standards (AERA, NCME, APA, 1999). Based on preliminary analysis of the pilot program, this framework seems a viable methodology to guide program success. Overall, the application of this framework was strait forward, and in general supported the validity of selection decisions made by residency directors who may use the new examination. As the administration expands to more provinces and larger samples of IMGs are tested, MCC will continue to explore our analysis and indications of reasonable use for this clinical evaluation.

This review of validity evidence has already proven useful for the design of the operational examinations in 2011. For example, the committee has reviewed a number of relevant OSCE examination blueprints and has chosen one that will be unique to the examination. Content is being donated by the partner organizations based on this blueprint in order to start a bank of stations. As well, new content is now being developed to address gaps in the blueprint. In terms of evidence consistent with response processes, there is considerable debate in the OSCE literature regarding the most appropriate scoring method for OSCEs; specifically should OSCEs be scored using a checklist or a rating scale. Research has been done to investigate the positives and negatives of both scoring approaches especially with regards to how raters complete each type of scoring instrument and what type of content is best suited to each type of scoring instrument. As summarized by Hawkins and Boulet (2008), checklists tend to be better for history taking and physical examination skills, any station in which specific actions are targeted, or if part of the intended purpose is to provide feedback to examinees. Rating scales tend to be better for judging complex aspects of clinical performance like communication skills or professional behaviors, or any station that does not have dichotomous tasks. That said, differences between scores assigned using the two methods tend to be small, reliabilities associated with each scoring instrument tend to be similar, and any issues that impact validity can be resolved with rater training. The NAC OSCE committee has reviewed this material on scoring as well

as other operational considerations and has decided to adopt a scoring approach that used ratings scales but will still provide checklist like scoring guidelines. Other statistical studies of the new program will revisit comparisons similar to that reported in this analysis of pilot data especially regarding the quality of the data, more in depth comparisons to scores on the MCCQE Part II, and a longer outcome analysis looking at sub scores from both examinations. It may be useful to explore the relationship between components of both test for examinees (e.g., history taking, communications or physical examination skills). Other future studies will include generalizability analyses to identify sources of error and variability in cases, raters and candidate performances.

Based on the pilot data presented here, evidence seems positive in support of this program as a useful and appropriate basis for selection in post-graduate training programs. This will serve to augment programs which already rely on a diverse and sometimes unique decision making process in each province and form a critical component of the pipeline that allows for competent care of the Canadian public.

Table 6. *Summary of the sources of validity evidence related to the 2010 NAC OSCE.*

Evidence	Description
Content Evidence	A new blueprint was created and steps were implemented to ensure content matches the blueprint. Some of the new content comes from the partner programs but steps to create new content are starting.
Response Process Evidence	Quality assurance steps were discussed. Reasons for choosing the scoring instrument were discussed.
Internal Structure Evidence	Item analyses were discussed, reliability of scores was described.
Relations to other variables	A comparison to MCCQE Part II was discussed.
Consequences	Survey program directors to track success rates of NAC candidates

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beckman, T.J., Cook, D.A., Mandrekar, J.N. (2005). What is the validity evidence for assessments of clinical teaching. *Journal of General Internal Medicine*, 20, 1159-1164.

Boulet, J.R., Smee, S.M., Dillon, G.F., Gimpel, J.R. (2009). The use of standardized patient assessments for certification and licensure decisions. *Simulation in Health Care, 4*, 35-42

Clauser, B.E., Margolis, M.J., Swanson, D.B. (2008). Issues of validity and reliability for assessments in medical education. In E.S. Holmboe and R.E. Hawkins (Eds.) *Practical guide to the evaluation of clinical competence*.(pp. 10-23). Philadelphia: Moseby Elsevier.

Downing S.M. and Haladyna, T.M (2009). Validity and its threats. In S.M. Downing and R. Yudkowsky (Eds.) *Assessment in health professions education*, (pp. 21–55). New York: Routledge.

Hawkins, R.E. & Boulet, J.R. (2008). Direct observation: Standardized patients. In E.S. Holmboe & R.E. Hawkins (Eds.) *Practice guide to the evaluation of clinical competence*. Philadelphia: Mosby Elsevier.

Hubley, A.M., Zumbo, B.D., (in press). Validity and the consequences of test interpretation and use. *Social Indicators Research.*

Kane, M. (2006). Validation. In (R.L. Brennan Ed.) *Educational Measurement.* (4th ed., pp 17-64). Westport CT: Praeger Publishers.

Margolis, M.J., Clauser, B.E., Winward, M., & Dillon, G.F. (2010). Validity evidence for USMLE examination cut scores: Results of a large-scale survey. *Academic Medicine, 85*, s93-s97.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp 13-104). New York: MacMillan.

Petrusa, E. (2002). Clinical performance assessments. In G.R Norman, C.P.M. van der Vleuten & D.I. Newble (Eds.) *International Handbook of Research in Medical Education*. (pp 647-672). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Swanson, D.B., Clauser, B.E., Case, S.M. (1994). Clinical skills assessment with standardized patients in high-stakes test: A framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education: Theory and Practice, 4*, 67-106

Van der Vleuten, C., Swanson, D.B. (1990). Assessment of clinical skills with standardized patients:

State of the art. *Teaching and Learning in Medicine, 2,* 58-76.