A CROSS-CULTURAL EXPLORATION OF PHYSICIAN ASSESSMENT

A CROSS-CULTURAL EXPLORATION OF PHYSICIAN ASSESSMENT

By AMITA MISIR, HON. B.ART SC., MD

A Thesis Submitted to the McMaster University School of Graduate Studies in Partial Fulfilment

of the requirements for

the Degree Master of Health Research Methodology

McMaster University

Department of Health Research Methods, Evidence and Impact

MASTER OF HEALTH RESEARCH METHODOLOGY (2020)

TITLE:

A Cross-cultural Exploration of Physician Assessment

AUTHOR:

Amita Misir, Hon. B. Arts Sc. (McMaster University), MD (Western University)

SUPERVISOR:

Dr. Sandra Monteiro, BSc, MSc, PhD

Assistant Professor

McMaster University

Department of Health Research Methods, Evidence and Impact

PAGES:

xi,150

LAY ABSTRACT

This is a case-study where the Objective Structured Clinical Examination (OSCE), a well-established, performance-based and resource-intensive Western medical education assessment tool, was introduced to the culturally different, resource-limited setting of Rwanda. What we wanted to evaluate is how the OSCE was received in the Rwandan medical training system.

What we found is that generally, the OSCE was received in a positive way. Both examiners and participants thought it was a relevant, realistic, feasible, valuable test for doctors in training. However, examiners also felt that the candidates did not do as well as they could have on the test not because they were fundamentally bad doctors, but because there were major gaps in their training. The OSCE therefore demonstrated its usefulness by identifying these deficiencies in training. Examiners felt that addressing these gaps in training was most important and should be done before any institutional body uses the OSCE results to decide who should get a medical license or not.

ABSTRACT

We conduct an evaluation of the cross-cultural 'export' of the Objective Structured Clinical Examination (OSCE), a well-established Western medical education assessment tool that is in keeping with Competency-Based Medical Education (CBME) principles, into the new socio-economic setting of Rwanda. The evaluation framework of 'assessment utility' is applied, where the utility of an assessment is described conceptually as the multiplicative function of its validity (V), reliability (R), educational impact (E), cost/feasibility (C) and acceptability (A). A mixed-methods approach of both quantitative and qualitative data analysis is used.

The quantitative findings support high content and face validity, high reliability, high acceptability and achievable cost and feasibility of the OSCE, all of which would suggest high utility. The analysis of qualitative data identifies some important threats to validity, namely perceived significant gaps in training in the internship program that were thought to likely be the underlying reason for the low mean assessment scores. This threat to the validity of the results appears to influence and limit the acceptability of the assessment in this context. While it is believed that it would be suitable as a formative assessment, primarily for the purpose of *'assessment for learning'*, it was not felt that it was currently acceptable as a summative or high-stakes *'assessment of learning'*, until and unless training deficits are addressed. Currently, the OSCE is seen to have greatest value in its potential for educational impact by acting as both a driver and a marker for change both at the individual and programmatic levels. Many principles of CBME and the concept of 'entrustability' as a criterion-referenced assessment standard were well-received cross-culturally, when training and assessment were viewed in tandem. Our study highlights the importance of using a comprehensive evaluation framework that includes both quantitative and qualitative methods to accurately characterize the utility of an assessment.

ACKNOWLEDGEMENTS

The author would like to convey her sincere gratitude and thanks to the following individuals for their support, time, efforts and expertise in conducting this research study:

Dr. David Cechetto B.Sc M.Sc M.Ed. Ph.D (Director, TSAM Rwanda-Canada)

Gisele Mukunde BA. (Research and Activities Assistant, TSAM Rwanda-Canada)

Ms. Kathy Johnston (Coordinator, IWK Simulation Program & TSAM Rwanda-Canada)

Dr. Ellena Andoniou (Project Manager, TSAM Rwanda-Canada)

Mrs. Aimee Utuza (Former Project Manager, TSAM Rwanda-Canada)

*Dr. Sandra Monteiro PhD (Scientist, McMaster Education Research, Innovation & Theory Program)

Dr. Stephen Rulisa MB, Mmed, PhD (Professor, University of Rwanda)

Dr. Emmanuel Musabeyezu MBChB (Education, Licensing and Registration, Rwanda Medical and Dental Council)

Dr. Lawrence Grierson PhD (Scientist, McMaster Education Research, Innovatoin & Theory)

Dr. Ellen Amster PhD (Associate Professor & Jason A. Hannah Chair in the History of Medicine, McMaster University)

Dr. Sandra Moll PhD (Associate Professor, School of Rehabilitation Science)

Dr. Cynthia Kenyon MD M.Ed. FRCPC (Team Leader, TSAM Rwanda-Canada)

Dr. Krista Helleman MD FRCPC (Team Member, TSAM Rwanda-Canada)

Dr. Kevin Coughlin MD MHSc FRCPC (Team Member, TSAM Rwanda-Canada)

Dr. Heather Scott MD FRCPC (Team Member, TSAM Rwanda-Canada)

Dr. Tharcisse Ngambe MBChB, FCPaeds (SA) (Team Member, TSAM Rwanda-Canada)

Dr. Anne Marie Tuyisenge MBChB (Team Member, TSAM Rwanda-Canada)

Dr. David Ntirushwa MBChB (Team Member, TSAM Rwanda-Canada)

Dr. Paulin Banguti MBChB (Team Member, TSAM Rwanda-Canada)

Mrs. Philomene Uwimana (Acting Director, Simulation and Clinical Skills Centre, College of Medicine and Health Sciences, Rwanda)

Special thank-you to my parents and husband, whose support has never faltered.

*Thesis supervisor

Table of Contents

List of Figures and Tables

List of All Abbreviations and Symbols

AAMC = American Association of Medical Colleges
ACGME – I = Accreditation Council for Graduate Medical Education - International
ACGME = Accreditation Council for Graduate Medical Education
AFMC = Association of Faculties of Medicine of Canada
AM = Amita Misir
AMA = American Medical Association
CBD = Competency By Design
CBME = Competency-based Medical Education
CFPC = College of Family Physicians of Canada
EPA = Entrustable Professional Activity
GAC = Global Affairs Canada
GDP = Gross Domestic Product
GMER = Global Minimum Essential Requirement
GMO = General Medical Officer
GNI = Gross National Income
GRS = Global Rating Score
HRM = Health Research Methodology
ID = Interpretive Description
IIME = Institute for International Medical Education
IO = Intraosseous
ITC = Item Total Correlation
MCQ = Multiple Choice Question
Mini CEX = Mini-clinical evaluation exercise
MNCH = Maternal, Newborn, Child Health
MOH = Ministry of Health
MSF = Multi Source Feedback
OSCE = Objective Structured Clinical Examination
PGME = Postgraduate medical education
PI = Principle Investigator
PMP = Patient Management Problems
PPH = Post partum hemorrhage
RCI = Royal College International
RCPSC = Royal College of Physicians and Surgeons of Canada
RMDC = Rwanda Medical and Dental Council
SAQ = Short Answer Question
SM = Sandra Monteiro
SP = Standardized Patient
TSAM = Training, Support, Access Model
VRECA = Validity, Reliability, Educational Impact, Cost/Feasibility, Acceptability
WFME = World Federation of Medical Education

Declaration of Academic Achievement

I, Amita Misir, declare this thesis to be my own work. I am the sole author of this document. No part of this work has been published or submitted for publication or for a higher degree at another institution.

To the best of my knowledge, the content of this document does not infringe on anyone's copyright.

My supervisor, Dr Sandra Monteiro, and the members of my supervisory committee, Dr. Lawrence Grierson and Dr. Ellen Amster, have provided guidance and support at all stages of this project. I completed all of the research work.

Chapter 1: Evolution of Medical Training and Assessment Globally

Paradigm Shifts in Medical Education

The education and training of medical physicians has undergone significant change in the last 150 years. The early 19th century saw great variability in approaches to medical education and program evaluation, particularly in the United States of America. The groundwork of the American Medical Association (AMA) Council on Medical Education and the Association of American Medical Colleges (AAMC) in the late 19th century and Dr. Abraham Flexner's landmark Flexner Report published in 1910 all served to propagate an evolution towards a standardized approach to medical education. The principles of this approach focused on pre-requisites for admission to medical school, a defined curriculum for medical trainees focused on amount of time spent in traditional science subject areas, access to hospitals and dispensaries where students should participate actively in patient care under supervision and salaried faculty who devote their time to teaching and research. (Barzansky, 2010)

Approximately 100 years later, we have two powerful models of standardized education that appear to be at odds with each other. One is *time-based* and directs attention to processes such as admissions and curriculum design. The other is *outcomes-based* and focuses more on the functional capabilities of the end-product (the graduate student, resident or practicing physician). An elegant metaphor contrasting these positions has been proposed as the 'tea steep' versus 'iDoc' model of medical training. The 'tea steep' or *time-based* model suggests that the right student (tea) is placed in medical training (hot water) for a fixed period of time. After a historically determined interval of time, we assume that a competent practitioner, like a good cup of tea, should result. The 'iDoc' or *outcomes-based* model suggests that medical schools and residencies, like factories, can be tailored to train doctors in specific functions adapted to user needs and desires.

The specifics of time and process that it takes to do so can be variable and is perhaps less important than ensuring that the desired outcomes are achieved.(David Hodges, 2010)

The increasingly popular *outcomes-based* model is also commonly referred as *competency-based medical education* (CBME). Below is a summary comparison of the two approaches (time-and-process based versus competency-based) as it relates to various aspects of medical education and training.

Table 1: A Comparison of the Structure- and Process-based vs Competency based Educational Programs§

A Comparison of the Elements of Structure- and Process-based Versus Competency-based Educational Programs

	Educational Program	
Variable	Structure- and Process-based	Competency-based
Driving force for curriculum	Content—knowledge acquisition	Outcome—knowledge application
Driving force for process	Teacher	Learner
Path of learning	Hierarchical (teacher ⇒ student)	Non-hierarchical (teacher ⇔ student)
Responsibility for content	Teacher	Student and teacher
Goal of educational encounter	Knowledge acquisition	Knowledge application
Typical assessment tool	Single subjective measure	Multiple objective measures (“evaluation portfolio”)
Assessment tool	Proxy	Authentic (mimics real tasks of profession)
Setting for evaluation	Removed (gestalt)	“In the trenches” (direct observation)
Evaluation	Norm-referenced	Criterion-referenced
Timing of assessment	Emphasis on summative	Emphasis on formative
Program completion	Fixed time	Variable time

§Reproduced with permission from (Caraccio, Wolfsthal, Englander, Ferentz, & Martin, 2002)

An Evolution in Assessment

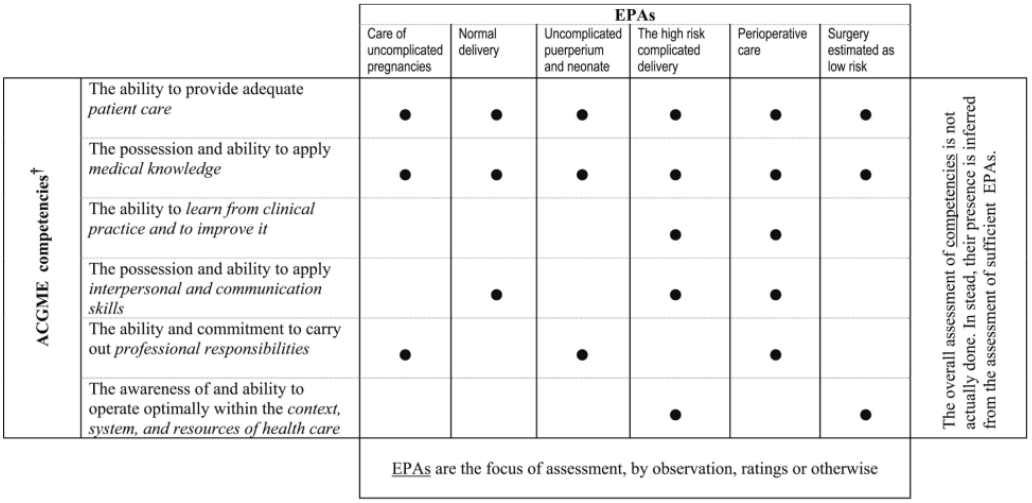
As medical education and training has evolved over decades, so too have the approaches to assessment of medical trainees. Van Der Vleuten states, ‘The historical development in competence assessment could be summarized as the continuous search for approximating professional…reality as close as possible while applying standardized test conditions’. (C.P. Van Der Vleuten, 1996) He describes the traditional view of competence as *‘trait-conception’* where competence is seen as an aggregate of different components or latent attributes (e.g., knowledge base, communication skills, attitudes) which were seen as relatively distinct from each other. (C.P.

Van Der Vleuten, 1996) The historical approach to assessment of competence has been focused on employing different methods of assessment (e.g., multiple choice questions, written simulations, learning process measures, live simulations) to isolate and measure these underlying 'traits' of competence. Development of competence was contemplated as being equal to the development in each of the component traits. (C.P. Van Der Vleuten, 1996)

In the CBME perspective, competence is not composed of underlying latent traits that we should be trying to measure using standardized tools. Rather, emphasis is on assessing the resulting integration of these traits; the integration of these competencies as manifested in the delivery of specific care activities. (Scheele & Ten Cate, 2007; C.P. Van Der Vleuten, 1996) Consequently, the focus of both *what we assess* and *how we choose to assess it* has changed.

From this arises the concept of "entrustable professional activities" or EPAs, which has become the focus of *what we assess* now. EPAs are defined as a unit of professional practice (task) that can be entrusted to a sufficiently competent learner (ten Cate, 2015). EPAs are important routine care activities that define a practice/specialty/subspecialty, are observable, executable within a time frame, and require an integration of competencies within and across domains to perform. (Englander & Carraccio, 2014; ten Cate, 2015) While competencies or 'traits' are person-descriptors (e.g., knowledge, skills, attitudes, values, abilities), EPAs are work-descriptors (e.g., discharge patient, counsel patient, lead family meeting, design treatment plan). The two concepts (competencies and EPAs) are not mutually exclusive and it is acknowledged that the concrete EPAs incorporate the conceptual competencies that are felt to be essential to professional practice. Figure 1 shows a conceptual map of this proposed relationship.

Figure 1. Conceptual relationship between EPAs and competencies*



*Reproduced with permission from the work of ten Cate & Scheele (Scheele & Ten Cate, 2007)

The shift in *how we assess* has not been so much in terms of the particular tools used (e.g., Multiple Choice Question - MCQ, Short Answer Question - SAQ, Objective Structured Clinical Examination - OSCE, mini-Clinical Examination Exercise - mini-CEX, long-case, incognito simulated patients, etc.) but rather a shift in the philosophy that drives how we choose to assess. In particular, there has been broad uptake on the criterion-referenced yet subjective concept of 'entrustability' in the assessment of medical trainees. Once again, this appears to parallel the general movement towards criterion-referenced (i.e., ready for practice or not) versus norm-referenced (i.e., below average/average/above average) assessment favoured in CBME (see Table 1 above). In the medical training setting, trust can be understood as "the reliance of a supervisor or medical team on a trainee to execute a given professional task correctly and on his or her willingness to ask for help when needed". (Ten Cate et al., 2016) The corollary to this is that being fully 'entrustable' refers to readiness to safely perform an activity/EPA without supervision. (Englander & Carraccio, 2014)

Generally the entrustability concept has been operationalized with a rating scale that includes pre-entrustable levels that defines varying levels of supervision required for the trainee

on a particular professional activity/EPA, up to fully entrustable without requiring supervision. It is proposed that trainees should be given increasing degrees of responsibility, and that decreasing degrees of supervision are required, as they move along a continuum from 'pre-entrustable' to 'entrustable'. An example of such a rating scale may be: 1 = trainee is able to be present and observe, 2 = trainee able to act with direct supervision, 3 = trainee able to act with indirect supervision, 4 = trainee able to act without supervision and 5 = trainee able to provide supervision. (O. Ten Cate, 2016; Ten Cate et al., 2016)

There are a few particular characteristics and consequences to note about using entrustability as a basis for assessment. Firstly, while it is acknowledged that ability (i.e., knowledge and skills) will be required to be considered entrustable, it is likely a necessary but not sufficient condition. Some key ancillary factors including integrity, reliability and humility likely also play into this decision. Generally these latter qualities are best seen over time, under real-life conditions, on multiple occasions and so the summative assessment of entrustability may be best suited to workplace-based assessment. (O. Ten Cate, 2016) Secondly, it is acknowledged that decisions of entrustment will likely have a substantial element of subjectivity as they may rely considerably on 'expert intuition'. Supervisors may differ greatly and different decisions will be made in similar situations. If entrustability assessments are employed in the real-world workplace setting as suggested, it also means that almost certainly the assessment conditions will not be standardized. However, this is no reason to abandon expert judgment or workplace-based assessment. (ten Cate, 2006) Indeed there is evidence that reliability is not conditional on objectivity or standardization. In recent years many studies have demonstrated that reliability can also be achieved with less standardized assessment situations and more subjective evaluations, provided the sampling is appropriate and sufficient. (C. P. van der Vleuten & Schuwirth, 2005)

Thirdly, the explicit linking of entrustability assessments to levels of supervision enables the assessments to acknowledge trainees' readiness for unsupervised practice of specified units of professional work. In doing so, it gradually enables them to become practitioners complying with the principles of outcomes/competency-based medical education: certification based on competence rather than on time in training. Lastly, entrustment decisions include patients in the equation and also the liability of the supervisor, linking assessment decisions more directly and obviously to patient and provider outcomes than do norm-referenced 'meets/does not meet expectations' scales. (Olle ten Cate, 2016)

Global Context

Much of the published literature about CBME and the parallel evolution in assessment frameworks has been dominated by institutions and authors from the relatively resource-rich settings of North America and Central/Western Europe. However, efforts for a shift from a traditional time-and-process to an outcomes or competency-based medical education has also seen some uptake in a more global level, both in resource-secure and resource-limited settings. Focusing on the postgraduate education phase of medical training, a recent article outlines the keen interest and uptake of the CBME-based training approach and associated institutional accreditation standards of the American-based Accreditation Council for Graduate Medical Education International (ACGME-I) in 15 sponsoring institutions in Asia, the Middle East and most recently in the limited-resource setting of the Caribbean nation of Haiti over the last 10 years. (Day & Nasca, 2019) Similarly, the Royal College of Physicians and Surgeons of Canada (RCPSC) that leads postgraduate education accreditation and standard-setting in Canada has established Royal College International (RCI) as its outreach platform. RCI currently has collaborative agreements with over 15 partner institutions in Asia, the Middle East and Eastern Europe to provide

consultancy services to strengthen postgraduate medical education globally. (Royal College of Physicians and Surgeons of Canada, 2020b)

The Challenges of Globalizing Medical Education

As outlined above, there is a global interest to translate the principles of CBME into medical education, not only in Europe and North America, where most of the published work on these concepts originates, but around the world in other resource-secure as well as resource-limited or emerging nations. International partnerships for this purpose can and most often do seem to be beneficial for both partners in this exchange. We should, however, consider the possible perils, particularly when this transfer-of-knowledge may be predominantly one-way with 'medical education products and services' moving primarily from resource-rich Western countries to settings that are often significantly different historically, culturally and economically.

Recent international medical education standards draw primarily on Western educational practices. Hays describes the most recent revised World Federation of Medical Education (WFME) standards as "…narrowing towards approaches more commonly seen in the developed world, particularly, Europe, a move that might limit both the achieveability and the relevance of the standards elsewhere." (Hays, 2014) While proponents of global standards acknowledge the need to respect local differences and celebrate diversity, they are at the same time promoting Western values, expressed in the language of 'core competencies' and the maintenance of equity through standardization. (Bleakley, Brice, & Bligh, 2008)

Postcolonial theory has been suggested to provide an important lens with which to examine medical education exports. (Bleakley et al., 2008; Whitehead, 2016) European colonialism was infused with the belief that colonizers took civilization and enlightenment with them. The superiority of European ideas and models was taken for granted, whether they were in the form of

religion, medicine, dentistry or legal and bureaucratic practices. (Whitehead, 2016) The Western medical curriculum, seen as an international text, is steeped in a particular set of cultural attitudes and rarely questioned. How can we be sure that modern global initiatives in medical education, which are largely advocated and funded by those in the 'modern, metropolitan West', who have the resources and influence to drive them through, are not just another type of 'domination by the advanced country over the developing nation'? (Bleakley et al., 2008)

The notions of 'validity', 'reliability' and 'generalizability' are highly prized in Western medicine and medical education. The more an education tool or process is deemed to be separable from a specific social, political, historic or cultural context, the more it is accorded value. (Whitehead, 2016) However, perhaps the answer to ensuring that the translation of Western educational processes and tools are a true cultural exchange, and not just another form of colonization, is first by recognizing that validity, reliability and generalizability are context dependent. Therefore, it would be prudent to establish their validity, reliability and generalizability in new global settings with as much rigor as was done in its original Western setting. Second, by recognizing that there are other relevant parameters – namely acceptability, cost/feasibility and educational impact – that deserve as much if not more attention when considering and evaluating these 'medical education exports'. Lastly, by creating opportunity to encourage truly bilateral flow of 'medical education products' and the evaluation of their implementation, so it is not all West-to-East or North-to-South, but also East-to-West and South-to-North.

When it comes to the 'exportation of educational products and services', which may broadly include accreditation, teaching curriculum and/or assessment processes and tools, there is relatively little to be found by way of both quantity and quality of research reporting on all these

aspects in the published literature. Of the reports available, often authors describe the plans for/the process of implementation, resulting performance score data and possibly the uptake of a new tool or process by an institution, leaving the reader to assume that it was therefore valid, reliable, affordable, feasible, acceptable and had the desired educational impact. (Al-Chalabi et al., 1983; Day & Nasca, 2019; Royal College of Physicians and Surgeons of Canada, 2020b; Schwarz, Wojtczak, & Stern, 2007; D. T. Stern et al., 2005; Stillman et al., 1997) When reporting on Western assessment tools in new global settings, there is often a focus on traditional psychometrics (i.e. demonstrating construct validity, factor analysis, cronbach's alpha, generalizability co-efficients) in published reports (Moiz, Ali, Rashid, Shariq, & Karim, 2019; Sasaki et al., 2005) While these reports are still certainly valuable and necessary – as noted, re-establishing psychometric data supporting validity and reliability is and should be part of the cross-cultural exchange process – it could be said that they may be leaving important gaps in what we need to be studying and reporting during these 'exports'.

Evaluating Assessment in a Global Setting

Van der Vleuten proposed the following conceptual equation to determine the utility of an assessment method for professional competence (C.P. Van Der Vleuten, 1996):

$$\text{Assessment Utility} = V * R * E * C * A$$

(V=validity, R= reliability, E =Educational impact, C=Cost and A=Acceptability)

Validity refers to the degree to which an assessment measures what it actually intends to measure. Reliability has been defined and measured in a variety of ways, however we will use the classical definition of reliability as the extent to which a measurement instrument can differentiate among individuals (Streiner, Norman, & Cairney, 2015). Educational impact refers to the ability of an assessment to influence the learning of the individual, or the curricular design of the learner

program for the institution. Cost may best be more broadly interpreted as feasibility, which would include both the monetary value and the ease of procuring the space, qualified staff, equipment, transportation and time needed to administer an assessment. Acceptability includes the entire belief system of people in relationship to assessment or an assessment method. (C.P. Van Der Vleuten, 1996)

The suggestion is that the utility of an assessment is the multiplicative function of these variables with different weights (w) associated with each of them. As there is no one perfect assessment, perfect utility is unobtainable. In practice, we will always need to compromise and assign different weights in different individual situations. For example, in a high-stakes examination with decisions having marked consequences on the future of examinees, reliability will probably have a heavier weight. In the context of formative in-training assessment, where the final decision may be based on many assessments, one may compromise reliability in favour of educational impact. What is important to note is that if any of the elements is zero, then the utility will be zero. A reliable, valid and low-cost test will have a short life if it's accepted by no one. (C.P. Van Der Vleuten, 1996)

Applying Van Der Vleuten's 'VRECA' conceptual utility equation is one way to approach the evaluation of assessment in global settings in order to ensure that important elements do not get missed when 'exporting' Western assessment methods. In the setting of introducing a Western assessment tool into a new global setting that may have both a different culture and different resource availability, it could be argued the variables of acceptability, cost/feasibility and educational impact are just as important and worthy of study as are the more commonly cited measures of validity and reliability. The former areas may not lend themselves well to exclusively

quantitative or numeric analysis. They are subjective concepts, unlike the often objectively or quantitatively-defined variables of validity and reliability.

Additionally, bias and equivalence are two pivotal concepts in the assessment of performance. Bias is said to occur if score differences on the indicators of a particular construct do not correspond to differences in the underlying trait or ability, but are rather attributable to incompatibilities of the underlying constructs, method or items of the assessment with respect to the sample or population being tested. Equivalence is usually accepted to mean the absence of bias. (van de Vijver & Tanzer, 2004) The use and adaptation of assessment instruments in light of this must be considered with respect to cultural validity and specificity in order to optimize the utility of such instruments.(Patel & Agius, 2017) van de Vijver et al. classifies and describes three types of bias (construct bias, method bias and item bias) as typical sources of bias in cross-cultural assessment. The presence or absence of such bias when attempting to 'transplant' an assessment derived in one cultural setting/population to another will not be easily identified by the simple reporting of standard quantitative/numeric measures of scores, correlations, factor analyses or reliability statistics. The potential for bias likely requires post-hoc reflection and discussion on both the process and the results of the administration of the assessment in the new cultural setting.

Thus, we propose that when 'exporting' a Western medical education assessment into a new cultural setting, both the "VRECA" framework for assessment utility and a mixed-methods research design should be applied to appropriately evaluate the activity. In addition to numeric, quantitative or statistical analyses aimed at 'objectively' measuring validity, reliability and other components of VRECA, generating qualitative data from open-ended reflection and discussion may prove to be an essential part of the picture, particularly when looking at the variables of validity, acceptability, cost/feasibility and educational impact. Both these quantitative and

qualitative methods are important to see if our findings from each analysis are generally convergent or divergent, and to gain a better understanding of why or how they may be so. We will apply this proposition to our particular case study of 'transplanting' the Western-originated OSCE into the new cultural setting of Rwanda and report on it in the following chapters.

In the remaining sections of this chapter, we set the scene by providing some contextual information about Rwanda, its medical training system, and how the idea of 'transplanting' the OSCE came to be. Then we will provide an overview of our study methodology, guided by a VRECA framework and mixed-methods approach. We will then be ready to move forward to subsequent chapters where the details of our evaluation study process and results will be reported and discussed.

Context of Assessment

Rwanda is a geographically small landlocked nation in sub-saharan Africa, with an area of 26,338 sq km (slightly smaller than the US state of Maryland). It has a total population of about 12.6 million people, giving it a population density that is among the highest in sub-saharan Africa. The capital and largest city is Kigali (about 1 million people). Official languages are Kinyarwanda, English, French and Swahili. The population is predominantly rural. Rwandans are drawn from just one cultural and linguistic group, the Banyarwanda, although within this group there are three subgroups: the Hutu, Tutsi and Twa. The people living in current day Rwanda were colonized by the Belgians in 1916. Belgians were the first to institute 'identity cards' in 1953 that labelled each individual as Tutsi, Hutu, Twa or Naturalized. Belgian rule systematically privileged the Tutsi as a superior race, leading to escalating tensions between the clans until independence from colonial rule in July 1962. Of note, Rwanda underwent a brutal genocide in 1994 during which an estimated 500,000-1,000,000 Tutsis and moderate Hutus were systematically killed in 100 days. The

country has been rebuilding itself physically, socially, politically and economically since that time. Clan identities have been abolished and replaced with a single identity of Rwandan. Christianity is the largest religion the country. The sovereign state of Rwanda currently has a presidential system of government and has had a single President, the Hon. Paul Kagame, for three consecutive terms since the year 2000.(Wikipedia contributors, 2020b)

Economically, Rwanda remains classified as a low-income economy as defined by the World Bank (Gross National Income or GNI per capita of $1,025 or less in 2018). For context, the GNI of high-income economies (i.e. Western nations like Canada, USA, United Kingdom, Netherlands) is classified at $12,376 or more by the same source/method. (The World Bank, 2020c) Rwanda's low-income economy history is nuanced however by a decimation of their Gross Domestic Product (GDP) around the time of the genocide, followed by a remarkable recovery. GDP growth rate was 35% immediately following the genocide and has been sustained at an average of 7.85% annual growth rate between 2000-2019. (The World Bank, 2020a) Much of this recovery has been indirectly supported/stimulated by annual net development assistance received, which is reported to be between $200 million - $1.1 billion US dollars (current US dollars) since 1994.(The World Bank, 2020b) Despite it low-income designation therefore, Rwanda is often praised as a model for positive, sustained economic growth and poverty reduction in the region, owing to its relative political stability, low corruption and strategic public investments in the roughly two decades following the genocide. (The World Bank, 2020d)

From a healthcare perspective, President Kagame has made healthcare an investment priority for the country since 2008 with significant success. In 2014/2015, health expenditure per capita for Rwanda was $53 US dollars (Republic of Rwanda Ministry of Health, 2020) with a total expenditure on health as a percentage of GDP at 7.53%. (World Health Organization, 2020)

Rwanda follows a universal health care model, which provides health insurance through a system called Mutuelles de Santé. The mutuelles are owned and managed at the level of Rwanda's thirty districts, and premiums of specific amounts are paid by citizens on a sliding scale ability-to-pay model, with the poorest citizens entitled to free health insurance and wealthiest paying premiums of US$8 per adult. There is a separate national health insurance scheme for public servants and soldiers. 90-95% of the population is insured. (Wikipedia contributors, 2020a) Similar to their socio-political-economic story, although key health indicators remain ranking low on a global scale both pre-and post-genocide, Rwanda has made enormous gains in the past two decades and now ranks relatively highly in its region (Africa) on several indicators including maternal mortality, infant and under-five mortality and life expectancy at birth. (Republic of Rwanda Ministry of Health, 2020; World Health Organization, 2020)

After a crippling loss of healthcare personnel due to both the genocide and HIV/AIDS, healthcare staffing in Rwanda also had to be rebuilt. Despite significant strides made in healthcare personnel education and training, Rwanda still falls well below the desired benchmarks for physicians, nurses and midwives in terms of quantity. Physician density was 0.064 per 1000 population and nursing and midwifery personnel was 0.832 per 1000 population (Republic of Rwanda Ministry of Health, 2020), well below the minimum 23 physicians, nurses and midwives per 10,000 population estimated by the World Health Organizaton (WHO) as being required to achieve adequate coverage.(World Health Organization, 2009) Creating local healthcare provider capacity in both quantity and quality continues to be a need and a goal for Rwanda.

Medical Training in Rwanda

From a physician perspective, Rwanda's oldest and largest School of Medicine and Pharmacy is run by the College of Medicine and Health Sciences, University of Rwanda. It has a

5-year (recently reduced from 6 years) undergraduate program, granting a Bachelor of Medicine and Surgery (MBChB). This program enrollment fluctuates, but generally admits 70-100 student per year. One or two smaller schools in Rwanda have also had some successful history in physician training. Upon successful completion of undergraduate training, medical graduates in Rwanda are then required to undergo a pre-licensure internship training period.

One year of internship is undertaken after graduating from medical school and before receiving independent licensure as a general medical officer (GMO). The design and implementation of the internship program has been delegated to the Rwanda Medical and Dental Council (RMDC), the country's national regulation and certification body for physicians, and falls under the Ministry of Health (MOH) rather than the Ministry of Education. During this pre-licensure period, intern physicians are deployed to district-level hospitals across the country where they are meant to gain additional experience/exposure to the major disciplines of internal medicine, surgery, obstetrics & gynecology and pediatrics while providing clinical service. At the end of their one year of internship, the intern physicians are generally granted independent licensure as a General Medical Officer (GMO) by the RMDC. The GMOs are then usually assigned to go on to give at least 2 years of service in the nation's district hospitals, thus forming the backbone of medical access and care for the majority of the Rwandan population and managing everything from low-acuity outpatient visits to emergency caesarian sections for complicated labour and delivery. Typically only after completing their 1-year internship and subsequent 2 years as an independently-practicing GMO do they become eligible to enter a specialist training program (where entry is based on a competitive application process).

At present, the assessment framework during internship is limited, based on a variably used clinical logbook and quarterly global evaluations. To date, there has been no set standards or

formal assessment program defined to determine fitness for independent practice and licensure at the end of internship. A logbook of clinical exposure is to be kept by intern doctors during their internship year, however this has been variably used and/or reviewed for adequacy. Even if it is used, it has limited comment on the competency of items logged as completed. Quarterly global evaluations are similarly variably used. The internship has thus far been mainly based on a structure-and-process model rather than a competency-verified and outcome-based system. As we shall see in the next section, the internship program recently came under some scrutiny, leading to a partnership with Canada to support a plan for quality improvement.

Partnering with Canadians in a Novel Assessment Exercise: Rationale for Execution, Evaluation and Study

In 2016, Western University in London, Ontario, Canada entered a five-year partnership project with the University of Rwanda under the Maternal, Newborn, Child Health (MNCH) program sponsored by Global Affairs Canada (GAC). The overall aim of this project is primarily to enable training and capacity-building that would strengthen MNCH care in Rwanda and thereby reduce maternal and child morbidity and mortality. The project is called Training, Support, Access Model (TSAM) in Maternal, Newborn and Child Health (MNCH) in Rwanda. At that time, the Rwanda Medical and Dental Council (RMDC) approached TSAM with concerns about an alarming and rising number of medico-legal cases that involved intern physicians and newly certified doctors. The majority of these involved perinatal care. The feeling was that there were significant gaps in the internship program and the request was for TSAM to provide some assistance in strengthening it. TSAM sponsored a stakeholders' meeting in which an internship strengthening plan was presented by Rwandans that would be based upon formulating and implementing a) internship site selection criteria b) formal mentorship of interns c) interns'

refresher course in MNCH and d) monitoring and evaluation. From the preceding plan, in addition to improving internship training site standards and sponsoring the development and delivery of a week-long Refresher Course in critical MNCH skills for all intern physicians, TSAM agreed to support the development and execution of a Rwandan' interns OSCE as part of the goal of monitoring and evaluation. For reference, the OSCE is a performance-based examination based on observation and scoring at a series of stations according to a set plan. Each station focuses on an area(s) of clinical competence and performance with a real patient, simulated patient, mannequin or patient investigations. (Harden, Lilley, & Patricio, 2016)

At this point, given the forgoing discussion, it should be recognized that we were proposing the 'transplant' of a distinctly Western medical education assessment tool (the OSCE) and quite probably its related concepts of 'competency', 'entrustable professional activities' and 'entrustability' into a place that was not only significantly different from a cultural perspective, but also from a resource-availability perspective. Support for the idea was certainly present from Rwandan partners, and human and financial resource provision would happen jointly with the support of both Rwandan and Canadian partners. However, viewing this through a 'post-colonial theory' lens as suggested by (Bleakley et al., 2008) and (Whitehead, 2016), it becomes important to be clear both on the rationale for the undertaking as well as to consider how we would evaluate its process and impact. Both aspects are described below.

The positive evidence supporting the validity and reliability of the OSCE as an assessment method has been well-documented in the literature and is broadly accepted (Harden et al., 2016; Norcini et al., 2011; C.P. Van Der Vleuten, 1996). However, reliability and validity are not inherent immutable traits of any instrument, it will be specific to the purpose, content, method (including preparation and sampling strategy) and context with which the instrument is developed

and administered. (Streiner et al., 2015; C. P. van der Vleuten & Schuwirth, 2005; C. P. van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010). It is therefore prudent to revisit these concepts even when deploying a 'validated and reliable' instrument in a markedly new and different setting where little literature exists or when significant innovations have been made to how the instrument was adapted and administered. With respect to cost/feasibility and acceptability, these are also context-dependent and little literature exists on what to expect on these fronts in a culturally different, limited-resource setting such as Rwanda. Alongside describing the development and implementation of the OSCE in Rwanda therefore, an evaluation of its validity, reliability, educational impact, cost/feasibility and acceptability per Van der Vleuten's framework (C.P. Van Der Vleuten, 1996) would be explored and discussed.

The question may arise that of all the assessment methods available (i.e. MCQ, SAQ, mini CEX, OSCE, etc.), why choose to introduce one that is traditionally one of the most resource-intensive into a limited-resource setting? The rationale for choosing the OSCE is both theoretical and practical and is summarized by the bulleted list below:

- The body of evidence supporting the reliability, validity, feasibility and acceptability of the OSCE over the past 30 years is extensive (Norcini et al., 2011), albeit mostly in Western well-resourced settings
- The administration of the OSCE is usually relatively centralized, which is easier to prepare, manage and control versus trying to implement a tool at 30+ internship sites across the country. This is particularly the case in the Rwandan context, where there is no clear pre-existing framework of faculty/supervisor appointments or faculty/supervisor development at those internship sites

- Via use of simulation, the OSCE allows for testing of emergency skills and invasive procedures that may be very relevant to the trainee's future practice and therefore to patient outcomes, but difficult to capture for assessment especially under direct observation during day-to-day practice
- In Miller's classic pyramid framework for clinical assessment, the OSCE sits at the 'shows how' level. (C. P. van der Vleuten et al., 2010) The pyramid classically has 4 tiers, beginning at the base with the assessment of 'Knows' (knowledge), then sequentially of 'Knows how' (competence/theoretical ability to apply knowledge), 'Shows how' (performance) and 'Does' (practice) in stacked tiers as it rises to the apex. (Miller, 1990) The 'shows how' and 'does' levels would appear to be best suited to the postgraduate level of trainee, as traditionally the 'knows' and 'knows how' levels are more relevant to and extensively tested at the pre-graduation level of medical training.
- In the context of a funded international partnership with medical expertise from another country, the appropriate resources and support to execute an initial proof-of-concept OSCE would potentially be more feasible than at other times

It is worth mentioning that in general, CBME assessment is increasingly emphasizing the importance and uptake of assessment at the 'does' level, with multiple objective measures, based on direct observation in the workplace, done over time with formative assessments leading to summative decisions for entrustability. (Caraccio et al., 2002; David Hodges, 2010) A single-point-in-time OSCE based on simulation, at first glance, does not necessarily appear to neatly match all these criteria. However, although compressed in time, a soundly designed OSCE will in fact rely on direct observation, use multiple objective measures (8+ stations) leading to a summative decision (an overall pass/fail decision for that OSCE). Furthermore, it should be noted

that while CBME certainly insists on assessment in the workplace, there is no one perfect assessment. Workplace-based assessment is subject to its own particular threats to validity and reliability; we should be cautious about relying on it as the only methodology of assessment to the exclusion of tried-and-true methods such as OSCEs and MCQs. Indeed, Van der Vleuten espouses the idea of moving away from trying to find the 'holy grail' of the perfect assessment and moving towards creating a programme of assessment that may include a variety of methods and methodologies, each with varying degrees of objectivity, structuring and standardization that may be from different levels of Miller's pyramid. (C. P. van der Vleuten & Schuwirth, 2005) Thus, the introduction of an OSCE to a limited-resource setting can still be relevant as one component of an assessment program that is part of the path to CBME.

Overview of Evaluation Study Methodology

As you will recall from previous sections of 'The Challenges of Globalizing Medical Education' and 'Evaluating Assessment in a Global Setting', we proposed that when 'exporting' a Western medical education assessment into a new cultural setting, both the "VRECA" framework for assessment utility and a mixed-methods research design should be applied to appropriately evaluate the activity. In our particular case of 'transplanting' the Western-originated OSCE into the new cultural setting of Rwanda, a mixed-methods design reporting on aspects of VRECA with data collected before, during and after the OSCE is applied. In Figure 2 below, the study methods are summarized and are also plotted on a timeline to indicate when they were executed. Additional details about each individual item in the methods summary is provided at the relevant times in the reporting of results in Chapters 2 & 3.

Figure 2: Summary of Study Methodology and Chronology

Study Methods Summary

a) Electronic survey to experts establishing EPAs and content validity of the OSCE (2017)

b) OSCE blueprinting, development and piloting (2017)

c) First iteration administration of OSCE with collection of scoring data (2017)

d) First iteration post-exam examiner paper surveys + in-person debriefs with field notes (2017)

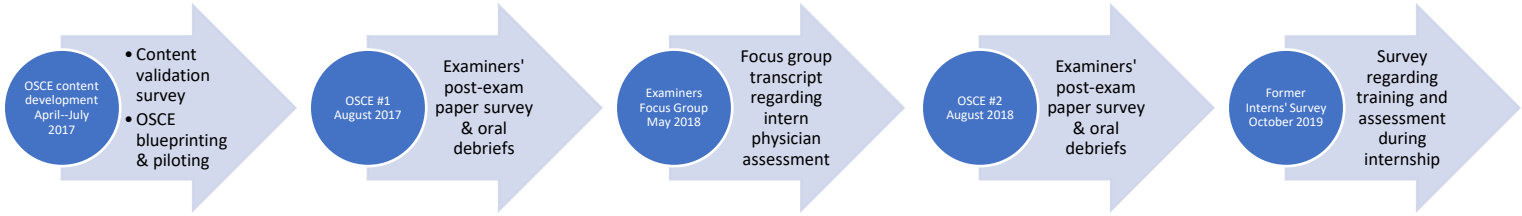
e) Examiners' Focus Group (2018)

f) Second iteration administration of OSCE with collection of scoring data (2018)

g) Second iteration post-exam examiner paper surveys + in-person debriefs with field notes (2018)

h) Electronic survey to all former interns who participated in OSCE/study to date (2019)

Study Timeline Summary



Data collected at the various points in the study was used to analyse the components of VRECA and to address potential for cross-cultural bias in the application of the assessment in the Rwandan setting. Chapter 2 takes a primarily quantitative analysis approach to this, using primarily descriptive and psychometric statistics related directly to the OSCE. Chapter 3 explores components of VRECA with respect to the OSCE and also more broadly explores intern physician training and assessment in general. A qualitative interpretive description analysis approach is taken in Chapter 3. Chapter 4 concludes with a review and synthesis of findings from both Chapters 2 and 3.

Preface to Chapter 2

In the previous chapter, we discussed some recent paradigm shifts in medical education towards the principles of CBME both in training and assessment. We described both the interest and the challenges in globalizing this paradigm shift, particularly with respect to the Western 'export' of 'medical education goods and services' to culturally and economically diverse settings in a post-colonial era. We proposed that some of the challenges could be mitigated by proper evaluation of Western 'exports'. Specifically in the setting of 'exporting' a Western-based assessment tool, the 'assessment utility' framework looking at all variables of validity, reliability, educational impact, cost/feasibility and acceptability (VRECA) should be used and explored ideally using a mixed-methods (quantitative and qualitative) approach. Applying this to the particular case study of Rwanda, we provided details on the specific context of Rwanda as a culturally and economically distinctly different setting from Canada, with whom it was partnering with in the 'export' of the OSCE for its intern physicians. We then gave an overview of the methodology of our evaluation study. In Chapter 2, we will rely primarily on quantitative data collected in the form of descriptive and psychometric statistics, as well as provide some relevant narrative details about our process of OSCE development and implementation, in order to evaluate the validity, reliability, cost/feasibility and acceptability of the OSCE in the culturally-different, limited-resource setting of Rwanda.

Chapter 2:
Development, Implementation and Evaluation of a Simulation-based, Multidisciplinary Objective Structured Clinical Examination (OSCE) in a Limited-Resource Setting

Introduction

Worldwide we have and continue to experience a paradigm shift from a structure- and process-based-based to competency-based medical education (CBME) and measurement of outcomes. While concepts of accountability and responsibility to the public for the practicing physicians has been a driving force behind the establishment of competency-based training for all physicians since the 20th century particularly in North America, the structure- and process-based system has still largely defined the training experience. Exposure to specific contents for specified periods of time (e.g., one month surgery, one month obstetrics, etc.) has been a mainstay behind physician training and certification. The paradigm shift from the current structure- and process-based curriculum to a competency-based curriculum and evaluation of outcomes has been called the Flexnarian revolution of the 21st century. (Caraccio et al., 2002) Instead of assuming vaguely-defined competence due to time/exposure spent during professional training, the emphasis is now on defining and demonstrating it in a real-world, practical manner.

This paradigm shift is documented in the international arena, in both undergraduate and postgraduate education. Focusing specifically on the assessment part of CBME, the literature provides reports of assessment programs and/or initiatives that embody the qualities of assessment espoused by CBME: multiple objective measures ("evaluation portfolio"), authenticity (mimics real tasks of profession), "In the trenches" (relies on direct observation), criterion-referenced and emphasis on formative assessments. (Caraccio et al., 2002)

Stern et al. reports on international collaborative work by the International Institute for Medical Education (IIME) to define the minimum essential competencies that all graduates world-wide should possess if they wish to be called a physician. Referred to as 'Global Minimum Essential Requirements' or 'GMERs', 60 competencies were defined across seven domains: (1) professional values, attitudes, behavior and ethics; (2) scientific foundations of medicine; (3) clinical skills; (4) communication skills; (5) population health and health systems; (6) management of information; and (7) critical thinking and research. Closely linked to the development of the GMERs, an assessment task force designed a 3-component assessment program with written, standardized and workplace performance assessments: a 150-item multiple choice question (MCQ) examination, 15-station OSCE and a 15-item longitudinal faculty observation form (the latter of which is meant to be administered at least 3 times over 3-6 months). Taken together, these components assessed a majority of the GMER competencies and were implemented by eight leading medical schools in China. (David T. Stern, Wojtczak, & Schwarz, 2003) The outcomes of this pilot implementation have since been reported. (Schwarz et al., 2007; D. T. Stern et al., 2005) Key findings support that an outcomes assessment process can complement traditional accreditation procedures and that instruments can be developed to assess GMER competencies appropriate for dissimilar cultures. Although validity, reliability and feasibility/cost data of the assessment tools were not explicitly published, evidence of acceptability and educational impact can be found. The GMER competence-and-assessment framework has been translated into multiple languages and thirteen countries have inquired as to whether the IIME would conduct pilot examinations in their countries. Sufficient credibility of the process and outcome was achieved for the Ministers of Education and Health of the People's Republic of China to approve publication of the actual performance data. The results of the GMER evaluation identified

strengths and areas for improvement of individual medical schools in relation to international standards, which has in turn re-invigorated medical curriculum reform in at least one participating medical school. (Xiao, Xian, Yu, & Wang, 2007)

Stillman et al. reports on three Chinese medical schools implementing a competency-based, outcomes-defined curriculum and assessment reform. This included the introduction of a new Standardized Patient (SP) program in both training and assessment. The assessment system was reformed to emphasize the importance of clinical ability. 50% of the final mark for the student was derived from a written examination and 50% from a clinical assessment using SPs. Standardized assessment tools based on direct observation of interviewing and physical exam skills were developed and used to measure student performance both before and after the curriculum-and-assessment intervention. One year after implementation, participating students significantly outperformed their counterparts who had been tested before the reform. The intervention has led to measurable improvement in students' clinical skills, in both increased performance scores and decreased variation among students. (Stillman et al., 1997)

In the context of postgraduate medical education (PGME) in global settings, the literature is limited but positive in the direction of CBME uptake. A recent article outlines the keen interest and uptake of the CBME-based training approach and associated institutional accreditation standards of the Accreditation Council for Graduate Medical Education International (ACGME-I) in 15 sponsoring institutions in the Middle East, Asia and most recently in the limited-resource setting of Haiti (in the Caribbean) over the last 10 years. (Day & Nasca, 2019) Similarly, the Royal College of Physicians and Surgeons of Canada (RCPSC) that leads postgraduate education accreditation and standard-setting in Canada has established Royal College International (RCI) as its outreach platform. RCI currently has collaborative agreements with over 15 partner institutions

in Asia, the middle East and Eastern Europe to provide consultancy services to strengthen postgraduate medical education globally. (Royal College of Physicians and Surgeons of Canada, 2020b)

A recent scoping review of PGME in sub-Saharan Africa revealed significant evolution in PGME training and programs between 1991 and 2016 (26 years). Among the challenges noted included 'the dearth of local PGME standards and lack of relevant curricula in some specialties' (Talib, Narayan, & Harrod, 2019) The internship year, which serves as a transition between medical school and either independent licensure and/or specialist residency in many countries, has been identified as a potential missed opportunity to improve quality and expand medical access in international settings, given that internship-trained generalist physicians provide the vast majority of patient care in many global settings where there is a lack of specialists. A recent publication reviews the experience of (and provides some guidance to) reforming the internship year as a roadmap for developing and emerging nations. A four-step process consisting of determining EPAs, identifying competencies and milestones associated with EPAs, designing instructional methods and developing assessment tools is outlined and its application to a local internship program in the United Arab Emirates (UAE) is described. (Bakir & Abdel-Razig, 2019)

As outlined above, it appears that there is an interest in and some precedent for applying principles of CBME in training and assessment in PGME not only in Europe and North America, where most of the published work on these concepts originates, but around the world in other resource-rich as well as resource-limited or emerging nations. One part of this approach includes developing and incorporating context-appropriate assessment methods that are consistent with CBME principles. Presently in Rwanda, graduates of the 5-year (previously 6-year) medical school program enter into a one year postgraduate internship overseen by the Ministry of Health (MOH).

The design and implementation of the internship program has been delegated to the Rwanda Medical and Dental Council (RMDC), the country's national regulation and certification body for physicians. During this pre-licensure period, intern physicians are deployed to district-level hospitals across the country where they are meant to gain additional experience/exposure to the major disciplines of internal medicine, surgery, obstetrics & gynecology and pediatrics. At the end of their one year of internship, the intern physicians are generally granted independent licensure as a General Medical Officer (GMO) by the RMDC. They are then typically assigned to a district hospital to serve as an independently-practicing GMO for the following two years before they are eligible to apply for specialist training programs.

To date, there has been no set standard or formal assessment program to determine fitness for independent licensure at the end of internship. A logbook of clinical exposure is to be kept by intern doctors during their internship year, however this has been variably used and/or reviewed for adequacy. The internship has thus far been mainly based on a structure-and-process model rather than competency-verified and outcome-based system. The present study reports on the process of developing, implementing and finally evaluating an 'exit OSCE' as a competency-based assessment tool for Rwandan intern doctors completing their internship year, who are about to embark on independent practice. In addition to describing the development of a standardized assessment tool that embodies many of the qualities of CBME assessment (multiple objective measures, authenticity, direct observation, criterion-referenced) in a novel context, report is made on the added challenges of executing this traditionally resource-intensive assessment method in a limited-resource setting while incorporating multi-disciplinary simulation-based examination content.

Assessment Background

In October 2016, a partnership was formed between the RMDC and TSAM (Training, Support, Access Model) MNCH (Maternal, Newborn, Child Health) Rwanda-Canada, with the goal of strengthening the Rwandan internship program. In addition to improving internship training site standards and sponsoring the development and delivery of a week-long Refresher Course in critical MNCH skills for all intern physicians, the need for proper monitoring and evaluation of interns and the internship program was identified. The development and execution of a Rwandan interns' OSCE was agreed upon as part of the goal of monitoring and evaluation. A research study was undertaken to assess the validity, reliability, feasibility and acceptability of the new assessment, as per the 'assessment utility' framework/approach for determining the utility of an assessment method for professional competence (C.P. Van Der Vleuten, 1996). Research ethics and institutional review board approvals were obtained from both University of Rwanda, College of Medical and Health Sciences, Rwanda and McMaster University, Canada to undertake the study. Two annual iterations of the same OSCE was run (2017 and 2018) at three provincial sites each year (one in each of Northern, Kigali and Southern provinces) with a target of 20 interns per site. Data was collected over about 3 years.

Methods

The following step-wise approach was developed as per recommended best practices for OSCE development (Harden et al., 2016) while incorporating CBME-based assessment principles (Caraccio et al., 2002). This resulted in the creation and administration of a direct-observation, simulation-based, inter-disciplinary, 10-station OSCE which included immediate feedback to the examinees. Following this, evaluation of the assessment was also undertaken using an 'assessment utility' framework focusing on validity, reliability, cost/feasibility and acceptability

of the assessment (C.P. Van Der Vleuten, 1996). The goal of the present study is to make observations on and evaluate the process of introducing a new approach to assessment and standard-setting in PGME in the context of a culturally different, limited-resource setting.

Exam Development and Materials

a. Creation of a 10-station OSCE blueprint
b. Writing of OSCE stations
c. Scoring tools and rubrics

Pilot/feasibility study

a. Piloting of OSCE stations with subsequent revisions
b. Resource planning and allocation

Participants and Recruitment

a. Recruitment and/or training of 10 specialist physician examiners, 15 simulation facilitators/standardized patients, 1 site co-ordinator, 1 simulation specialist and 1 local OSCE physician lead per site, per year for each of 3 provincial OSCE sites
b. Recruitment and orientation of target 20 intern physicians per site, per year at each of 3 provincial sites

OSCE Administration/Data Collection

a. OSCE administration and scoring data collection
b. Self-administered examiner surveys conducted immediately after the examination
c. Electronic survey to intern physicians including questions about the OSCE assessment (conducted 1-2 years after finishing internship)

Psychometric Analysis

a. Performance measures – deriving descriptive statistics from participant scores, calculating Cronbach's alpha and item-total correlations to evaluate reliability.

b. Examiner Surveys – Descriptive statistics

c. Intern Physician Surveys – Descriptive statistics

The results of the above step-wise process is presented in this chapter, documenting the utility of the assessment as well as the unique challenges in the process of introducing a new approach to assessment and standard-setting in PGME in the context of a culturally different, limited-resource setting.

A more in-depth exploration of the meaning and potential impact of this OSCE was planned using a follow-up discussion for a sample of examiners one year after the first exam administration. An examiners' focus group was held with the objective of reviewing exam results with them, exploring in greater depth the validity, acceptability and potential educational impact of the OSCE and more broadly exploring their attitudes towards intern physician assessment and standard setting. The results of the focus group are presented in Chapter 3.

Results

Exam Development and Materials

a. Creation of a 10-station OSCE blueprint

To ensure content validity in developing the OSCE blueprint to the final form displayed in Table 2, and to help identify potential specific stations for the OSCE, a survey was circulated to 9 Rwandan physician experts spanning the four specialities (Paediatrics, Obstetrics and Gynecology, Internal Medicine, Surgery) as well as general medicine. These experts also have experience in a variety of leadership roles in education (i.e., Dean of Rwandan medical school), regulation (i.e., Chair of Education committee for Rwandan Medical and Dental Council) and practice (i.e., recently-graduated physicians who have served at District Hospitals within the last 5 years). 7 of the 9 invited respondents replied.

17 items for ‘key performance domains’ (i.e., Y-axis on the OSCE blueprint/grid in Table 2) were initially proposed, comprised of ‘entrustable professional activities’ (EPAs). EPAs are defined as important routine care behaviours and activities of a physician that are able to be judged as ‘entrustable’, which is defined as readiness to safely perform the activity without supervision. By definition EPAs should be observable and measurable and thus may be most suitable for OSCE-based assessment. The proposed EPAs were largely based on the Association of American Medical Colleges (AAMC) EPAs (Association of American Medical Colleges, 2014) as well as a few additional items gleaned from Good Medical Practice (General Medical Council, 2013) and the Rwandan Medical Intern Handbook (Republic of Rwanda MInistry of Health, 2011). All 17 proposed key performance domains ranked as either ‘quite relevant’ or ‘very relevant’ by >75% of respondents (a somewhat arbitrary but reasonable threshold for establishing validity) (Streiner et al., 2015), supporting content validity for these items. These 17 EPAs were reduced to 13 to be included in the OSCE blueprint, with the rationale that some EPAs needed to be dropped either because they were not optimal for testing in an OSCE setting and/or overall station/time restrictions of the OSCE due to practical considerations. Content validation for the general ‘Areas of Professional Practice’ (i.e., X-axis) for the blueprint was not specifically requested. This was because looking at both the training objectives and the future immediate practice of interns, it was felt to be fairly obvious that the X-axis should include the four major areas of professional practice as identified, namely obstetrics & gynecology, paediatrics, internal medicine and surgery.

In the survey, experts were then provided with a comprehensive list of skills/problems outlined under each specialty in the interns’ logbook (which is part of the Rwandan Medical Intern Handbook) that could reasonably be tested in an OSCE. Certain items that were considered too difficult or impossible to test in realistic way in an OSCE setting (e.g., reduction of a fractured

bone) were not included. The experts were then asked to indicate the priority level of each possible skill/problem/objective in terms of its inclusion for the basis of an OSCE station for a time-and-station limited OSCE. Using the somewhat arbitrary criterion where a skill/problem should be rated as 'Very High priority' or 'High Priority' by >75% of respondents, a number of skills/problems were eliminated as putative OSCE stations: 5 from paediatrics, 6 from obstetrics, 20 from surgery, and 22 from internal medicine. After the above process, more than 50 skills/problems still remained.

At this point in time, key decisions regarding number of stations feasible for the OSCE needed to be considered. In order to maximize the number of examinees and avoid exhaustion, it was felt the exam should not go beyond 3 hours to allow for 2 administrations per day (one in morning, one in afternoon). Cost of specialist examiners was a significant budget consideration that would motivate towards fewer stations; however this needed to be balanced against best practices evidence that suggests a minimum of 10-12 stations lends sufficient reliability to the test (Harden et al., 2016). To ensure that stations remained realistic to practice and were not so brief that the exam felt reduced to 'monkeys doing tricks' (Barman, 2005), at least 10 minutes per station was allotted. To provide some formative value to trainees taking the exam, a brief post-station feedback time was to be incorporated, which would mean less time for exam station testing. The combination of the above resulted in the decision for a ten-station OSCE.

Given that time/resources would permit only 10 stations, either certain skills/problems would need to be combined into one station and/or some further decisions as to which skills/problems to include as OSCE stations would need to be made. Certain other questions on the experts' survey, including relative importance of urgent/emergent vs elective problems, as well as feasibility considerations during pilot testing, helped to guide such decisions. It was proposed

that since there will be 10 stations, each specialty area should be allocated 2-3 stations. Within this, where appropriate, each of the three stations should be chosen on the basis of one emergency/acute scenario, one scenario incorporating a procedure and one routine/ward scenario. The resulting final blueprint with stations 1-10 demonstrating content distribution is shown in Table 2.

Table 2: OSCE Blueprint

Key Performance Domain (Entrustable Professional Activities or EPAs)	Areas of Professional Practice			
	Obstetrics & Gynecology	Paediatrics	Internal Medicine	Surgery
EPA 1 - History & Physical	4,7,10	3,6	9	8
EPA 2 - History/Physical Interpretation & Differential Diagnosis	4,10	1,3,6	9	8
EPA 3 - Plan of Investigation	10		5,9	8
EPA 4 – Interpret/communicate test results	7,10	6	9	
EPA 5 - Management Plan	4,7,10	1,3, 6	9	2,8
EPA 7 - Handover in transitions of care		6		
EPA 8 - Urgent/emergency care	4	1		
EPA 11 - General procedures	4	3	5	2
EPA 12 – Educate/counsel patients	7			
EPA 13 – Recognize limitations		6		

Note: Numbers in blueprint grid represent stations that cover those areas of Key Performance and Professional Practice. Certain EPAs were dropped (from original 17 to final 10 shown above) as they were not felt to be well-suited for OSCE assessment.

b. Writing of OSCE stations

Following the blueprinting and selection of stations for the OSCE, station development was undertaken which included writing, reviewing and revising of the stations. This was done in collaboration with Rwandan and Canadian physician specialists to combine experience/familiarity with station writing (Canadians) with relevance to local context (Rwandans). During station development, a number of features were considered and integrated to enhance realism, with the goal of maximizing the face validity and credibility of the examination (Harden et al., 2016). These key features are outlined in Table 3.

Table 3. Key Features of the developed OSCE Model

Features	Description
• Integrated	All stations were designed to integrate at least two (or more) EPAs, as would be expected in most real patient care situations.
• Multidisciplinary	Three of the emergency/acute stations included 'role players' of nurse, midwife and/or another physician. The candidate was expected to work within a team with these individuals, as they would in their future posts as General Medical Officers (GMOs) in district hospitals.
• Simulation-based	All stations utilized either standardized patients (5 stations) and/or mannequins/task trainers (5 stations) to allow assessors to observe first-hand how assessments and/or procedures were conducted by candidates.
• Entrustability-assessed	An 'entrustability' global rating score (GRS) was used in conjunction with 'prompt checklists' to be used as an aid for examiners when selecting a global rating. (See more details in Descriptive Results section)
• Resource-appropriate	Checklists and materials provided were developed/adapted such that they would fall within the scope of availability or expectation of the local setting.

c. Scoring Tools and Rubrics

For each scored item on the OSCE, the same global rating scale was used. The 6-point scale is based on 'entrustability' and was defined as follows:

A (1)	B (2)	C (3)	D (4)	E (5)	F (6)
NOT ENTRUSTABLE For independent practice	BORDERLINE NOT ENTRUSTABLE For independent practice	APPROACHES ENTRUSTABLE For independent practice	ENTRUSTABLE For independent practice	PROFICIENT For independent practice	EXPERT For independent practice

Entrustability was defined as 'readiness to safely perform the activity without supervision' (Englander & Carraccio, 2014; O. Ten Cate, 2016). The OSCE examination was timed as an exit-exercise at the end of the intern training year and the candidates were about the embark on independent practice. It would therefore be reasonable to expect candidates to be performing at a level that is entrustable for independent practice (i.e., score of 4), although an a priori 'pass' standard was not set for this exercise. Each station had a variable number of EPA subscores plus an Overall Global Rating Score (GRS), giving a total of n=51 scored items across the 10 stations.

While station development including guideline checklist-items of key features to look for when scoring was created collaboratively, the scoring scale itself was pre-determined by the Canadian OSCE lead (Amita Misir, AM) and based on the 'entrustability' concept. The entrustability definition as well as the scoring scale as outlined above was presented and discussed with examiners during their examiner training sessions (see 'Recruitment and/or Training of Examiners and OSCE Staff' below for details of training). Additional supplemental reading material about EPAs and entrustability from the Association of Faculties of Medicine of Canada (AFMC) was also provided for examiner reference. (Association of Faculties of Medicine of Canada, 2016)

The above process of establishing partners and experts willing to support the initiative, OSCE blueprinting, and station development and review took place over approximately 1 year. Enabling factors included keen interest from the RMDC for the idea and locating several specialist physicians in each discipline from both Canada and Rwanda willing to help with station creation, mostly from their goodwill on a volunteer basis, with some direction from the OSCE lead (AM). Because the Western University partnership with Rwanda had existed for over a decade, pre-existing familiarity and trust from many of those relationships likely helped with early buy-in. Timing was also key in that RMDC had heard growing concerns over the clinical practice of some of their recently-trained intern physicians and TSAM had just recently sponsored a large, multi-stakeholder meeting regarding the internship program at the request of RMDC. As a result, they were likely eager to engage in any effort that potentially provided data about and/or a solution to address what they saw as challenges with their internship program. Also, as TSAM was known to have funding and at least a 4-5 year time horizon for its activities, this also likely helped to inspire trust and motivation. Personal meetings between the OSCE lead (AM) and the TSAM Project Director ensuring that the initiative was in line with overall TSAM goals and objectives, that it

could and would be reasonably funded and that 10 stations could be accommodated was also critical.

Pilot/Feasibility Study

Once the OSCE blueprint and stations had been developed and drafted, piloting of the stations was the next essential step in determining feasibility. Piloting of all 10 planned OSCE stations was undertaken over several sessions, with certified specialists acting as examiners. The 'examinees' during the pilot were either specialty-specific postgraduate trainees or certified specialists assuming the role of examinee. The process of piloting resulted in several important revisions of the stations and a better recognition of what resources would be required. For example, a few stations were felt to be too long and were therefore modified/simplified to limit and focus the objectives and tasks. Certain local materials (e.g. local chicken leg for intraosseous (IO) access procedure) were found to be either difficult to procure and/or challenging to successfully utilize. For the IO procedure, it was felt that a mannequin task trainer might prove more reliable than a chicken leg and 18 gauge straight or spinal needles (which is what would be available locally) would be made available at the station even though IO needles would make the procedure easier.

As real patient-care areas would be best suited for the exam, it was decided it would be held on weekends or evenings at each site to minimize interference with patient care. 3 administrations of the exam were held over 3 weeks at 3 different sites (North, South, Kigali) on weekends and/or evenings. Each site had one sitting in morning (10 candidates circulating through 10 stations), one sitting in afternoon (10 candidates circulating through 10 stations) or the same was held over two consecutive evenings. A list of required resource and their allocation can be found in Table 4.

Table 4. Resource requirements and allocation

Resource required	Allocation
I. Human Resources	
▪ OSCE lead	Canadian TSAM volunteer (all sites)
▪ Local OSCE lead	Rwandan physician specialist, also serving as examiner (one per site)
▪ Simulation specialist	Canadian TSAM volunteer
▪ Local Simulation specialist	Rwanda healthcare simulation specialist
▪ Administrator/Data manager	Rwandan TSAM staff
▪ Examiners	Rwandan physician specialists recruited by RMDC (10/site)
▪ OSCE Site Co-ordinator	Local nurse/midwives recruited by TSAM (1/site)
▪ Simulation facilitators/actors	Local nurses/midwives recruited by site co-ordinator (7/site)
▪ Standardized patients	Local nurses/midwives recruited by site co-ordinator (8/site)
II. Physical Space	
South	On hospital grounds, divided between Internal Medicine outpatient building (5 clinic rooms with common hallway) and Obstetrics & Gynecology outpatients (5 clinic rooms around atrium). Also central training room for examiner/facilitator training and reception, small medical records room for examinee reception/orientation.
North	On hospital grounds, divided between Internal Medicine outpatient building (5 clinic rooms in close proximity) and Gender-Based Violence Centre (5 clinic rooms in separate building). Also large central training room for examiner/facilitator training and reception, smaller technology room for examinee reception/orientation.
Kigali	On hospital grounds, in large Outpatient Clinic area. Up to 15 clinic rooms available with common corridor. Hospital simulation centre and division meeting room used for examiner facilitator/training. Examinees received and oriented at simulation centre, then escorted to Outpatient Clinic for exam.
III. Equipment	
▪ Mannequins	Borrowed and transported from University of Rwanda nursing schools or simulation centres
▪ Furniture	Used existing furniture in clinics/on hospitals sites
▪ Supplies (i.e. needles, sutures, needle drivers)	Purchased through local hospital pharmacy (supplies) and/or borrowed from hospital (hardware)
▪ Bifold Clipboards	Purchased locally at papeterie
▪ Timers	Purchased through Amazon, brought to Rwanda
▪ Whistles (in place of buzzers)	Purchased in Canada, brought to Rwanda
IV. Financial resources	
▪ For provision of examiners, facilitators and SP stipends, catering, printing, supplies, transport stipends for all training and exam sessions	Provided by TSAM for a total cost of: 11,188,800 RWF ($13,281.89 USD) in 2017 = $245.96 USD per examinee (54 total examinees) 10,863,997 RWF ($12,548.94 USD) in 2018 = $261.44 USD per examinee (48 total examinees)

Participants and Recruitment

a. Recruitment and/or training of examiners and OSCE staff

For each year of the exam, 10 specialist physician examiners, 15 simulation facilitators/standardized patients, 1 site co-ordinator, 1 simulation specialist and 1 local OSCE physician lead for each of 3 provincial OSCE sites (target 20 interns at each site) were recruited and (in most cases) trained. Recruitment of this quantity of human resources can be a daunting task in any setting. This Rwandan context proved no different and presented its own unique challenges. For example, practicing specialist physicians were recruited as examiners, which required training (2-hour evening session) and work (10-hour weekend/weeknight) that was not in their 'job description' and was on top of their regular clinical duties (vs faculty members who may have this as part of their academic duties). Recruitment was facilitated largely by RMDC leadership contacting their members to serve as examiners, often with a great deal of help from 3-5 key RMDC/TSAM physician specialists (in pediatrics, obstetrics and internal medicine) who had been involved/invested in the OSCE from the start on a mostly volunteer basis, as well as help from TSAM administrative staff . It was also aided by being able to provide some level of stipend for their time, although this was not advertised at the time of recruitment but given only after the examination. However, there was widespread dissatisfaction from examiners with the quantity/type of stipend distributed in the first year, so slightly more funds were allocated somewhat differently the following year (i.e., less money for the mandatory training, but more money on the actual exam day for examiners). Although the hope had been to get local examiners at each site, this was not always possible particularly with the limited availability of specialists in the country. They would then need to be brought in from a neighbouring province, adding travel and accommodation logistics and costs. There was also the occasional incident of an examiner cancellation at the last-minute or being called for urgent patient care during the exam. One such

incident resulted in a non-trained francophone examiner stand-in who failed to understand the scoring concept and did not provide any scores for the 20 interns at that station. Occasionally we were fortunate to have a trained 'back-up' physician on-site to temporarily stand in for missing examiners. Having such an option beforehand can add a cost but can be very valuable when needed.

As a practical assessment exercise of this magnitude had not been administered by the RMDC before, there was no existing administrative support to facilitate its execution. A site co-ordinator had to be recruited at each site, usually a nurse or midwife. This person was approached by TSAM, and again pre-existing relationships with these individuals through other TSAM activities (such as existing TSAM mentorship programs at these hospitals) facilitated this. This individual then became instrumental in securing the physical space, staff and supplies required for their site, and again a small stipend was provided. There was no existing medical school SP program that was employed to provide SPs, so 8 SPs needed to be recruited ad hoc at hospitals by the site co-ordinator and then trained. As our OSCE was heavily simulation-based, having one simulation specialist on site to help manage and support stations which required a lot of technical equipment proved essential for set-up, trouble-shooting and take-down of those stations. In addition to this simulation specialist, 7 simulation facilitators/technicians per site (usually nurses/midwives) were recruited then trained to run the actual equipment at the relevant exam stations. All local SPs/facilitators were provided modest stipends for attending training and for exam participation.

Training for the examiners, SPs and simulation facilitators took place at each site, usually 1-2 days before the actual exam administration. While the exam itself was booked on weekend/evening time, getting the right timing and permission for training was important to minimize impact on clinical service. Generally SP and technician training was conducted from

approximately 11 am - 2 pm, followed by examiner training from 5-7 pm on the same day. SPs and simulation facilitators were assigned their specific roles, then divided into two groups: the SPs were individually trained/practiced in their roles by the OSCE lead (AM) and the simulation facilitators were trained in their roles by the simulation specialist. As the trainers were often English-speaking only and the nursing/midwifery staff often had limited English, having the OSCE site co-ordinators present to assist with interpretation for the SPs/facilitators proved critical to utility of the training.

Examiners were oriented to the overall exam, assigned stations and introduced to scoring rubric and approach. The entrustability definition as well as the scoring scale (as outlined previously) was presented and discussed with examiners during their examiner training sessions. Additional supplemental reading material about EPAs and entrustability from the Association of Faculties of Medicine of Canada (AFMC) was also provided for examiner reference. (Association of Faculties of Medicine of Canada, 2016) Mock stations were held and examiners would score the performance independently, then discuss their scores in attempt to build some consensus around scoring. Finding spaces that would be adequate for all the equipment and people involved in training was sometimes a challenge, but in most cases a training room(s) or meeting space(s) on the hospital site could be found and/or re-purposed for this.

b. Recruitment and orientation of intern physicians at each site

A total of 104 Rwandan intern physicians participated in the OSCE, with n=55 participating in year 1 (2017) and n= 49 participating in year 2 (2018). The target had been 20 per site (or 60 per year), however a number of challenges prevented full participation.

RMDC had been requested to assign the intern physicians to each site, based mainly on geography. Enabling the intern physicians to participate in the exam meant not only did they

require release from active essential clinical duties but also required transport and potentially accommodation to the exam site, the latter of which was supported financially by TSAM. Formal communication from the RMDC went out to all internship sites that intern physicians were to be excused from clinical duties to attend the exam, however it was often on very late notice which made achieving this difficult. The other important piece was contacting and confirming each candidate to ensure they could get to and participate in the exam. As the OSCE was technically a no-stakes exercise with no pass/fail, no possible impact on their progression, they were encouraged and supported to participate but would generally not face any consequences if they did not. The RMDC enlisted the help of the interns' own leaders (who are interns themselves) to communicate and organize themselves by site. Despite best efforts by intern leaders, this proved to result in a lot of challenges with timely commitment, transport and attendance. Even among the candidates that did participate, several were very late due to transport or other reasons causing delays in exam start and/or missing parts of the exam.

If this were to be conducted in the future, RMDC should consider making participation (if not pass/fail status) a clear requirement of completing internship and therefore of getting an independent license. Official correspondence should flow directly from the RMDC to each internship site as well as each intern citing required exam dates and logistics well before the exam. Interns should be excused by at least 3 pm the day preceding the exam until 3 pm the day following the exam to allow for travel. If possible, exam dates that do not conflict with Umuganda (the country's national service day, last Saturday of every month) should be chosen or letters of permission clearly excusing them from Umuganda duties should be provided. This may decrease many last-minute challenges and frustrations related to trying to get candidates to exam sites and

increase likelihood of full participation, which should be a goal particularly for such a resource-intensive examination.

Once at the exam site, all interns were oriented to the study as well as the OSCE exam. Consents were obtained for the study and a brief primer/tips on what to expect during the OSCE and how to maximize success was provided. In addition to powerpoint slides/verbal briefing, in most situations, a visual/tactile orientation was given on the infant and pelvis mannequins that they would see at some of the exam stations. This process generally took 30-60 minutes before the start of each examination.

OSCE Administration/Data Collection

a. OSCE administration and scoring data collection

The actual OSCE exam was generally administered on a weekend day at a provincial large hospital site to avoid conflicts with clinical care and physical space, with one sitting in the morning (8 am-noon) and one sitting in the afternoon (1 pm-5 pm). In one instance, two consecutive evenings were done instead, which proved difficult for set-up and take-down of stations in between clinic days. The weekend schedule generally worked out well, however, several issues were noteworthy in the process:

i. Timelines – as a general theme, this was a challenge throughout the entire planning process: choosing training and exam dates, approaching clinical directors of hospital exam sites for permission to use their facilities and staff, informing clinical directors of hospitals to excuse interns for the exercise, confirming & making arrangements for the participation of examiners and interns in particular. None of this was generally done with sufficient advance notice. The importance of starting exam sessions on time was also not respected. This often led to unnecessary stress, cold reception from hospital leaders, excessively long

exam days and suboptimal participation in the exam, at times bordering on the need for exam cancellation days before a scheduled exam. This is in part due to cultural norms of last-minute activity, newness of the exercise without an appreciation for the necessity of advance planning and timeliness, and at times unclear and/or unreliable ownership and accountability from key stakeholders and identified leaders to get things done in a timely manner.

ii. Politics and local resource allocation – When undertaking the OSCE exercise and this study, initial partners included RMDC as well as medical school clinical faculty members who were leaders in postgraduate medication education, simulation and research. This was both a strategic and natural progression, as RMDC was in charge of the internship program while the medical school faculty had better supply of and access to resources such as experienced clinical instructors and simulation centres with equipment, space and specialists/technicians. As plans progressed, the model of holding regional exams at provincial hospitals using local physicians as examiners evolved and was preferred by RMDC for a number of legitimate reasons including cost, logistics, local empowerment and future sustainability. The consequence of this was that most of any supporting funds and training would flow directly to peripheral exam sites and their OSCE examiners/staff recruited by RMDC and TSAM, rather than to the central university hospitals or their departments/faculty.

At this point in time there was a marked change in the level of support from medical school and its faculty members. In particular, access to simulation equipment and specialists became a major challenge. Peripheral sites did not have these resources and it did not appear that the medical school was willing to loan these supplies/services any

longer. In the weeks leading up to the first iteration of the OSCE, there was a serious scramble to find a work around to this situation. Thankfully and again facilitated by pre-existing relationships with TSAM, critical simulation equipment was procured on loan from local nursing school simulation centres. Some equipment that had been previously purchased by TSAM for their mentorship program was borrowed and certain equipment was transported from Canada by Canadian volunteers. After the first successful iteration of the OSCE, in the second year, the medical school faculty members seemed again more willing to rejoin the initiative with renewed support and enthusiasm including access to their simulation equipment and personnel, which did make things significantly easier. Going forward, relationship-building and a memorandum of understanding particularly between the Ministry of Health/RMDC and the Ministry of Education/medical school may be good to clarify common goals and commitments related not only to the OSCE but also to the internship training program in general.

iii. Equipment/staff transportation – Given the model of distributed provincial examination sites, equipment and staff transportation was of critical importance. Rwanda is a small country and in recent years, has invested in major roadways. Still, its hilly landscape and relatively limited major roadway network meant up to 3 hour travel-times in what could be challenging conditions, particularly in evenings/nights. The existing infrastructure and staff of TSAM provided this critical piece in this case. Wood was purchased and backboards were hand-cut to enable safe transport of expensive loaned mannequin equipment. A TSAM-owned large truck with a TSAM-supported driver was used to pick up equipment from various sites. This same truck was ultimately loaded and packed with all necessary equipment, materials and core TSAM staff/volunteers to travel to each exam

site. A secured room on hospital site was requested to unload and store all equipment for the duration of training and exams. All of this logistics and transport costs is not reflected in the cost-breakdown provided in Table 2. In the future, plans would have to be made to somehow provide this procurement and transport. Using local OSCE leaders to procure equipment from local nursing and medical school simulation centres near each OSCE exam site may be possible and would minimize transportation logistics/requirements.

Scoring was done by examiners, on paper, during the exam. After finding missing information in scoresheets after the first exam sitting, a double-check of scoring data was introduced. At the end of each exam administration, the paper scoresheets were collected by TSAM staff member(s) and checked individually for completion. If gaps were found, it was immediately brought to the attention of the examiner to correct. The data from the scoresheets was inputted by the TSAM staff member into an SPSS format database.

Psychometric Analysis

A total of 104 Rwandan intern physicians participated in the OSCE, with n=55 participating in year 1 (2017) and n= 49 participating in year 2 (2018). One individual had to be excused from the exam early due to illness, this subject was dropped from the dataset for a total of n=103 subjects for analysis. Descriptive statistics, reliability statistics and item-total correlations were conducted to both summarize the performance of the examinees and explore the performance properties of the test itself.

a. Performance Measures

The OSCE was comprised of a total of ten stations. Each station included i) a variable number of EPA subscores (depending on the number of EPAs tested at that station) and ii) a single

Overall Global Rating Score (GRS) for performance at that station as a whole. Taken together, this gave a total of 51 scored items that could be used for assessment and analysis.

Most score sheets were checked and verified for completeness following the exam. For items where scores were missing for a particular candidate, the score was replaced by the mean score of their exam cohort (i.e., 2017 or 2018) for that item. Generally speaking, for 48/51 total scored items, <3% of data was missing. (i.e. 0-3 out of a total possible n=103 candidate scores were missing for any given scored item). The most notable replacements/substitutes to the dataset were as follows:

a) 20 candidates had no scores for 3 scored items at Station 2 in 2017 (due to the oversight of a last-minute stand-in examiner who had not attended the examiner training sessions). As outlined above on how missing data was handled, these scores were replaced by the mean score of their exam cohort for those items.

b) For one station subscore (Station 6, EPA 7), the doctor was to initiate a phone call to a specialist for further assistance/transfer with their patient and the candidate was to be scored on the quality of their communication. If the candidate did not think to initiate the call, they were automatically given a score of 1 ('Not entrustable').

c) 1 candidate completed only 50% of the stations due to a delay in arrival.

Descriptive Statistics

The mean scores for the overall exam and for each station are listed in Table 5 and Table 6, respectively. Both EPA subscores and Overall Global Rating Scores are reported separately and in combined fashion as they represent different measures of the same underlying construct i.e. a competent physician.

Table 5: Descriptive statistics for OSCE scores - overall

Scoring items included	Mean (n=103 subjects)	Standard Deviation
All EPA subscores from all 10 stations (n=41 items) Excludes Overall Global Rating score (GRS) from each station	3.64	0.36
All Overall Global Rating Scores (GRS) from all 10 stations (n=10 items) Excludes subscores from stations	3.56	0.42
All subscores + all overall Global Rating Scores (GRS) from all 10 stations (n=51 items)	3.63	0.37

Table 6: Descriptive statistics for OSCE scores – by station

Station	Mean of station subscores only, (SD)	Mean of station Global Rating Score (GRS) only (SD)	Mean of subscores + GRS, (SD)
Station 1 Pediatrics	3.16 (0.85)	3.12 (0.97)	3.15 (0.86)
Station 2 Surgery	3.59 (0.74)	3.52 (0.82)	3.56 (0.74)
Station 3 Pediatrics	3.42 (0.43)	3.28 (0.56)	3.40 (0.42)
Station 4 Obstetrics/Gynecology	3.57 (0.78)	3.57 (0.97)	3.57 (0.79)
Station 5 Internal Medicine	3.14 (0.75)	3.27 (0.78)	3.18 (0.73)
Station 6 Pediatrics	3.71 (0.52)	3.69 (0.65)	3.71 (0.53)
Station 7 Obstetrics/Gynecology	3.72 (0.80)	3.70 (0.89)	3.71 (0.81)
Station 8 Surgery	3.80 (0.87)	3.80 (1.09)	3.80 (0.89)
Station 9 Internal Medicine	3.90 (0.61)	3.79 (0.69)	3.88 (0.60)
Station 10 Obstetrics/Gynecology	3.93 (0.70)	3.87 (1.01)	3.92 (0.74)

Reliability Statistics

A Cronbach's alpha was calculated to assess the reliability of the test administrations. The specific formula used to calculate this was as follows:

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

Where:

- N = the number of items.
- $\bar{c}$ = average covariance between item-pairs.
- $\bar{v}$ = average variance.

This was calculated based on a few different proposed methods of overall test-scoring, to explore how choosing to score the test in different ways may affect its reliability. The results are shown in Table 7 below.

Table 7: Reliability statistics for OSCE scores (2017 & 2018, all candidates)

Scoring items included	**Cronbach's Alpha**	**N of items**
All subscores from all 10 stations Excludes Overall Global Rating score (GRS) stations	0.87	41
All Overall Global Rating Scores (GRS) from all 10 stations Excludes subscores from stations	0.647	10
All subscores + all overall Global Rating Scores (GRS) from all 10 stations	0.90	51

Item-Total Correlations

Item total correlations (ITCs) were calculated to explore how each individual station performed. The total was calculated as the mean of all subscores + GRS scores of all stations. The item was calculated as the mean of all subscores + GRS for a given station. The results are given in Table 8 below.

Table 8: Item-Total Correlations by Station (2017 & 2018, all candidates)

Item (defined as mean of all scores at station)	**Corrected Item-Total Correlation**
Station 1 - Mean of EPAs + GRS scores	.199
Station 2 - Mean of EPAs + GRS scores	.275
Station 3 - Mean of EPAs + GRS scores	.461
Station 4 - Mean of EPAs + GRS scores	.464
Station 5 - Mean of EPAs + GRS scores	.240
Station 6 - Mean of EPAs + GRS scores	.636
Station 7 - Mean of EPAs + GRS scores	.329
Station 8 - Mean of EPAs + GRS scores	.314
Station 9 - Mean of EPAs + GRS scores	.275
Station 10 - Mean of EPAs + GRS scores	.269

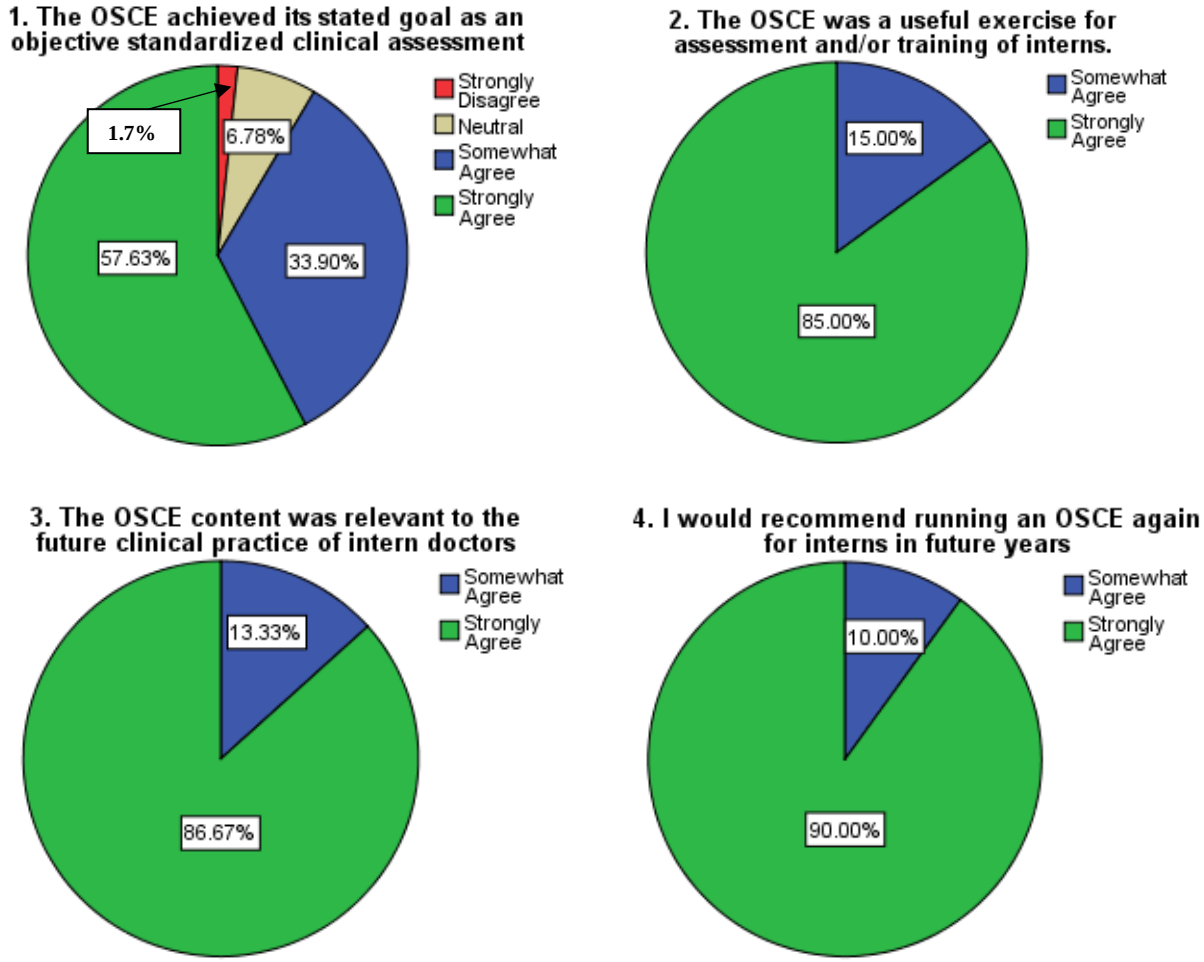
b. Examiners' Survey

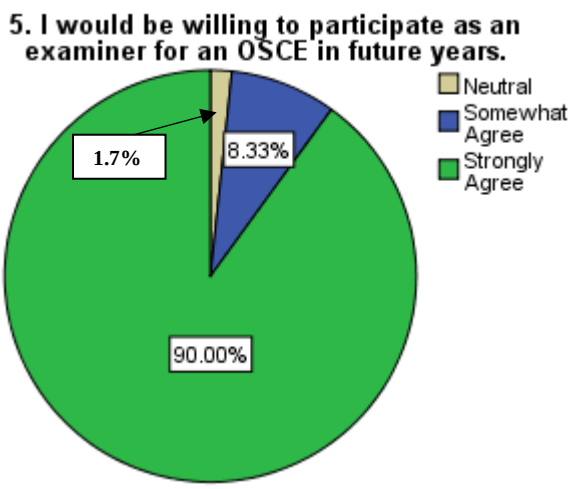
Immediately following each OSCE, participating examiners were asked to complete a brief self-administered anonymous survey on paper. The purpose of the survey was to assess content validity, face validity and acceptability of the OSCE post examination, from the perspective of the examiners. Every examiner at each site completed the survey in both years, for a total of n=60. Descriptive statistics of the survey are summarized in Figure 3.

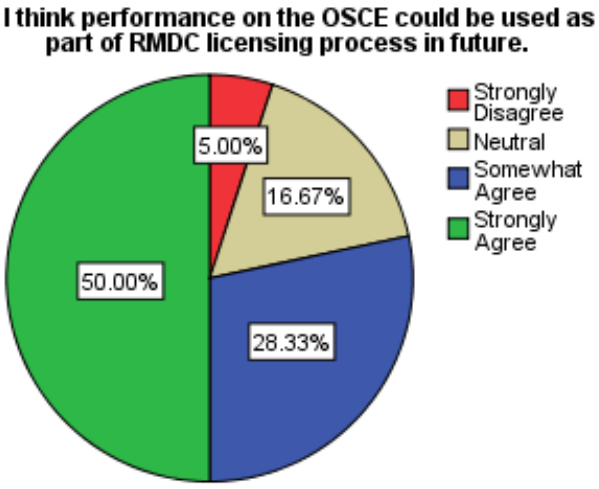
Figure 3: Examiners Post-OSCE questionnaire results

Part 1 – Questions regarding the Interns' OSCE overall

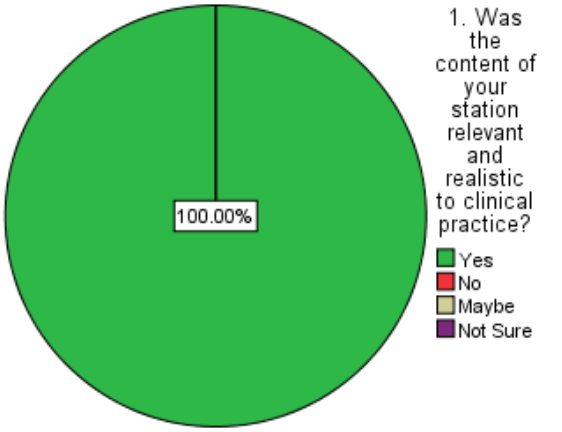
Available response options same for all questions in Part 1:
Strongly Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Strongly Agree

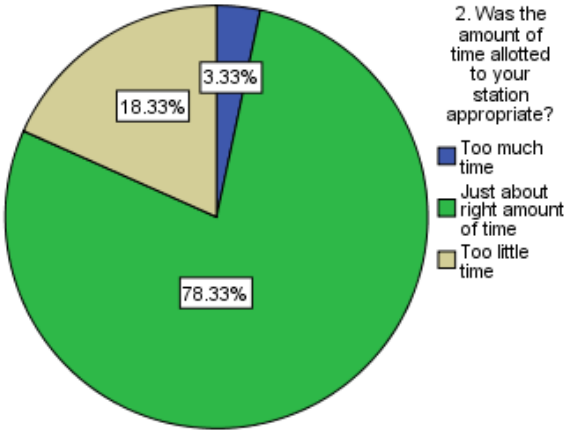


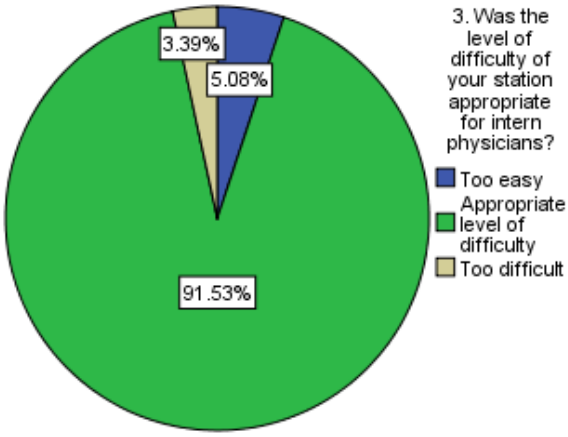


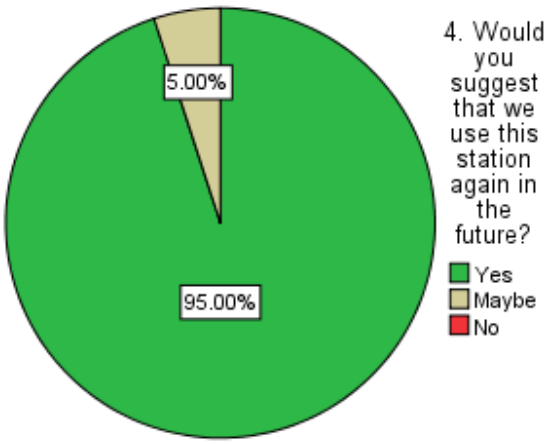


Part 2 – Questions regarding the OSCE station where you were an examiner today









c. Intern Physicians Survey

A survey was sent electronically to all interns who participated in the study. The content of the survey covered several aspects of the internship experience (i.e., hospital facilities, program governance, clinical exposure and supervision, continuing professional development, feedback and evaluation). The survey was sent in October 2019 (approximately 1-2 years after the study subjects would have finished their internship) using a Qualtrics platform to all 104 subjects. Four email reminders for completion were sent every 2-14 days over 2 months with one additional opportunity to complete the survey at 3 months post initial distribution. A total of 61 interns responded to the survey (58.7% response rate). Three questions from the survey specifically related to the OSCE were included and their descriptive statistics are summarized in Figure 3.

Table 9: Intern Physicians Survey results (n=61 respondents)

Survey item (related to the OSCE)	**Somewhat to Strongly Disagree (%)**	**Neither agree nor disagree (%)**	**Somewhat to Strongly Agree (%)**
I felt the OSCE was a useful learning experience.	9	5	86
I felt the OSCE was a good method of assessment.	15	7	78
I think that an OSCE exercise should be run for future interns.	4	6	90

Discussion

To our knowledge, this is the second report in the peer-reviewed English-language literature of the development, implementation and evaluation of a simulation-based OSCE for postgraduate physicians in a limited-resource setting. It is the only one that systematically collects, analyzes and reports on descriptive and psychometric data in relation to its validity, feasibility/cost, acceptability and reliability in such a setting. The other publication is a brief report on the pilot

implementation of a pre-admission OSCE for admission to a family medicine training program in Haiti. It describes some similar challenges in implementation with respect to resources and notes local support at education leadership levels (i.e., Graduate Medical Education Committee) to continue it in that setting. (Sainterant, Clisbee, & Julceus, 2019)

The literature reporting on OSCEs for medical trainees in limited resource settings is scant. The publications that do exist are in the context of introducing the OSCE as part of a formal university undergraduate/pre-graduation medical curriculum supported by university faculty (Abdelaziz, Hany, Atwa, Talaat, & Hosny, 2016; De Almeida Troncon, 2004; Vargas et al., 2007). While some of these experiences share comparable characteristics, there are some unique circumstances introduced in this Rwandan case. First, the administration/organization was undertaken primarily in partnership with a medical licensing body (the RMDC) rather than a university. Second, it was executed at a critical time-point for the intern physicians, that is on the cusp of when they were to receive their independent license to practice without supervision. These two factors mean that such an OSCE could potentially be used as a high-stakes licensing examination, emphasizing the need for convincing evidence of validity and appropriate reliability. Third, the lack of affiliation with a formal medical faculty and timing of the exam being after the postgraduate physicians (examinees) were already deployed across the country at their internship sites meant that certain costs, resources and logistics were particularly challenging, as described throughout the results section. Despite these unique challenges and many others, our results would suggest that such an OSCE can be executed with reasonably good evidence of validity, feasibility, reliability, and acceptability.

Validity

Both the rigorous methods by which the OSCE blueprint was developed and the post-OSCE examiners' questionnaire results support the content and face validity of the examination. 100% of examiners agreed that the OSCE content was relevant to the future clinical practice of intern doctors and that the content of their station was relevant and realistic to clinical practice (Figure 3). In addition to careful blueprinting, writing, piloting and revision of the 10-station OSCE, we believe that incorporating the characteristics listed in Table 3 (Integrated, Multidisciplinary, Simulation-based, Entrustability-assessed, Resource-appropriate) helped contribute to this validity. It is probably true that incorporation of some of these features increases both the cost and complexity of the examination in terms of the resources and training required (Table 4). However, forgoing these elements runs the risk of making the OSCE prone to some of the potential weaknesses pointed out by others. It may become limited in its ability to measure what the individual would do in real-life situations, measure parts in isolation that is not equivalent to measuring the whole integrated performance and not assess the in-depth knowledge and skills that are necessary for postgraduate students (Barman, 2005; De Almeida Troncon, 2004).

Reliability

In this OSCE, stations 1-5 were centered around simulation using mannequins or models for emergent situations or surgical/practical procedures. Stations 6-10 were centered around interviewing and management skill using standardized patients (SPs). Given the relative novelty and complexity of simulation-based stations 1-5 in particular, it is interesting to note how these stations 'performed' relative to others in the OSCE. Table 8 displays the item-total correlations that was used to do this. Station 1 had the lowest correlation at 0.199, while stations 2-5 had a correlation co-efficient between 0.24-0.464 and stations 6-10 had a co-efficients of 0.275 - 0.636.

The usual rule of thumb is that a good test item should correlate with the total score between 0.3 – 0.7. If it is below 0.3 it is presumed that it is potentially not accurately measuring the same construct as the rest of the test (in this case, physician competence) and if it is great than 0.70, it is likely a redundant item that could be removed. (Streiner et al., 2015) Using these guidelines, it would seem that stations 2-10 fell in or near this 'acceptable' range, including the 'novel' simulation-based stations. The reasons behind why Station 1 appeared to have a low correlation to the total score warrants further reflection before determining whether or how this station could be changed or eliminated.

The overall reliability the OSCE was calculated at Cronbach's α of 0.90 if all available scored items were utilized in the final score (Table 7). It is likely that several factors contributed to this relatively moderate-to-high reliability index. First, fairly involved examiner training was undertaken. The three-hour training included introduction to the OSCE and its purpose, the content of the ten stations, how to apply the checklists and scoring scheme and importantly, time for practice-scoring where multiple examiners from the same discipline would score a 'mock candidate', then discuss similarities/differences in their scores. Second, a single, simple scoring scale based on the relatively novel but purportedly intuitive concept of 'entrustability' (ten Cate, 2006) was applied. It was constructed to anchor upon the gradations that an examiner may reasonably be expected to see and differentiate between. Third, a total of 51 items were made available for scoring and as more of these items are incorporated into calculating the final score, the greater the reliability of the test (see Table 6). All these are suggested as legitimate ways to enhance the reliability coefficient (Streiner et al., 2015). In addition, a standard 30 minute pre-OSCE orientation performance was given to all examinees outlining what to expect for the exam, what would be expected of them, mannequin use and how to optimize their performance. This

was felt to be particularly important given the novel nature of the assessment. All simulation facilitators and SPs were also given a 3-hour training session.

Optimizing Validity and Reliability

The optimal way in which to score the examination would arguably be both valid and maximally reliable. To this end, it is suggested that incorporating all EPA subscores + the GRS from each station appears to achieve this and gives an overall reliability index (Cronbach's α) of 0.90. The validity of this approach is grounded in the fact that each subscore at each station is based on a separate EPA and is therefore scoring a different aspect of physician performance in that clinical context. The GRS, while one would expect might reflect the subscores, was not predicated upon them. It was a gestalt on the overall performance at that station and not specific to any EPA.

Although checklists were provided as guides to observation, it was explicitly emphasized that ultimately it was examiner *judgment* – not a strictly applied checklist and/or tally-count – that would determine the score for any given item. The goal of this hybrid approach (checklist-guided, global rating scale based scoring) was to strike a balance between the threats to validity that strict checklist scoring can introduce and the enhanced reliability that checklists may provide (Harden et al., 2016). Strict checklist scoring has been criticized being overly reductionist in approach (Harden et al., 2016) and punishing true expertise where not every step is undertaken but rather the right ones at the right times (Barman, 2005). Using global rating scales can help overcome these threats to validity (Harden et al., 2016).

It has been suggested that for high-stakes examinations, a reliability index such as a Cronbach's α or g-coefficient of greater than 0.7 or 0.8 is necessary (Harden et al., 2016). However, one should also bear in mind the limitations of reliability metrics and the relative

paramount importance of validity. Central to this is that reliability indices such as the Cronbach's alpha (based on classical test theory) or g-coefficient (based on generalizability theory) aim to describe how much of the observed variance in a given sample comes from true between-subject differences vs random-error (i.e. raters, items, sites, test execution, etc.), as suggested by the below general form of the reliability equation (Streiner et al., 2015):

$$Reliability = \frac{Subject\ variability}{Subject\ variability + Measurement\ error}$$

It is important to note then that if the true baseline between-subject variability is low, as may be the case for highly trained physicians, the reliability index will remain low regardless of accurate measurement. Equally important to consider is the relevance of reliability to the assessment goal. If the goal of the assessment is to determine if candidate A is better or different then candidate B, then an instrument that has a high ability to discriminate between subjects becomes important. If however the main goal is to determine whether or not candidate A and candidate B meet a certain standard, as is increasingly the case in CBME, then the ability to discriminate between candidates – which is the reliability metric in the sense of being able to isolate and measure between-subject differences – becomes less relevant. What becomes more relevant is convincing evidence that your assessment measures what it intends to measure (i.e. validity) and that appropriate standard-setting is undertaken. For these reasons, it has been acknowledged that validity should not be sacrificed for reliability and feasibility (Harden et al., 2016) and that a test with high reliability is not always better than a test with a lower reliability (Schuwirth & van der Vleuten, 2011a).

Cost/Feasibility

The approximate cost of the OSCE as reported in Table 4 was about 11 million RWF ($12,000 USD) per year, which worked out to about $250 USD per examinee. This was inclusive of all examiner and facilitator/SP training sessions. This appears to be on the lower end of the wide

cost-range ($11 to $1200 USD per candidate) for OSCEs as reported in the literature (Harden et al., 2016), but still not insignificant in the context of the Rwandan limited-resource setting. This figure is best seen as a rough estimate of cost. There is a wide variation in the way that costs are calculated and reported in the literature, often resulting in a 'high-end cost' and 'low-end cost' (Reznick, Smee, Baumber, Blackmore, & Berard, 1993). Our study was no different with respect to this variation. It is important to mention some elements that in this case may contribute to either a 'high-end' or 'low-end' reporting. Most notably, the reported cost did not include any cost/payment for three key OSCE leadership personnel: OSCE lead, Simulation specialist, Administrator/Data manager. These costs are difficult to estimate for a number of reasons: their specific hours were not systematically tracked, their hypothetical local hourly rate or payscales are not defined, the demands of their roles were variable year-to-year (often more involved in first-iteration) and their actual cost would probably change based on whether or not these were stand-alone jobs (unlikely) versus responsibilities integrated into a larger job description (more likely). It also did not include costs of the actual vehicle that helped with transport of materials between sites (as this was a TSAM project owned vehicle). It also did not include the costs associated with development and piloting of the OSCE stations. Conversely, the reported cost did include all costs related to the many training sessions for the OSCE. If this was to become a routine year-upon-year practice, then with sufficient years of institutional memory, such extensive training may no longer be required.

Acceptability

In terms of acceptability, Figure 3 suggests that the OSCE was perceived as a useful, relevant and realistic learning/assessment tool by the significant majority of examiners. A clear majority felt it should be run again in future years and would be willing to be an examiner again.

Examiners generally felt their station content was of appropriate difficulty level, time allotment and relevance. This is important because besides the validity, reliability, cost and feasibility of an assessment tool, the acceptability likely strongly influences is uptake and use going forward (C.P. Van Der Vleuten, 1996). Table 9 similarly supports the acceptability of the OSCE from the examinees' perspective, with greater than 90% of intern physicians indicating it was a useful learning experience that should be run for future interns, and nearly 80% endorsing it as a method of assessment.

It is interesting to note that the question with the greatest distribution of opinion was around whether the OSCE could be used as part of a licensing process or standard in the future. About 20% of the respondents were either neutral or somewhat disagreed with this statement. Informally during debriefing sessions held with examiners immediately after the exam, it seemed that many of the examiners recognized that the average performance of the intern doctors fell below the 'entrustable for independent practice' rating (see Table 5 and Table 6). There was some concern that if the OSCE was implemented as a standard that had to be 'passed' at present, many doctors potentially may not 'pass' and there would be even fewer doctors to serve patients in an already resource-strapped setting. It is possible this, as well as other factors, could explain some of the reservation in using the OSCE as a licensing standard-setting assessment tool. This and other perspectives on intern doctors' assessment was explored in a follow-up focus group, the results of which will be analyzed and reported in chapter 3.

Conclusion

Before closing this analysis, a few limitations of the study warrant mention. One limitation is that while strong evidence of content and face validity has been presented, other evidence that would contribute towards validity would have strengthened the case. It would have been

interesting to explore concurrent or criterion-related validity by comparing OSCE scores to other assessments of competence. However, there was little by the way of existing assessment data that could be considered suitable for comparison. The only existing assessment was the intern logbook; as mentioned this was not rigorously or consistently used. Similarly, exploring for evidence of predictive validity by comparing OSCE scores of exam candidates to some measure of their future physician performance could also be a helpful contribution, however no such future measures/assessments exist.

Another limitation to note is that the 'E' or Educational Impact component of the 'VRECA' equation was not explicitly measured or discussed here. This is mostly due to the limited time-horizon/follow-up period to the study, study scope/resource limitations as well as the 'exit-timed' nature of the OSCE which occurred at the end of internship for individual interns. Educational impact, defined as ability of an assessment to influence the learning of the individual, or the curricular design of the learner program for the institution, is likely best measured and tracked over time. At the individual learner level in the context of a training program, to determine meaningful educational impact (vs short-term knowledge increase), an assessment is probably best done while in midst of training with some follow-up measures (self-reported and/or performance based) that would suggest a change (hopefully positive) in the individual's learning due to that assessment, presumably at some later time-point in the training program. While the former interns self-reported that they found the OSCE to be a useful learning experience (see Table 9) which would be suggestive of positive educational impact at the individual level, they were not specifically asked or tested on the quality or quantity of their learning as a result of the OSCE assessment. At a programmatic level, the timeline of educational impact is likely longer. Assessment results first have been analyzed, synthesized and presented to key program

stakeholders and their implications discussed. Following this, the intention and practicality of changing curricular design would need to be explored. Finally, the ultimate design, implementation and evaluation of any such programmatic changes would need to be tracked.

In summary, this study was an exercise in introducing Western-derived rigorous standardized performance-based assessment tool (the OSCE) in a culturally different, limited-resource setting for postgraduate physicians. We have presented evidence of validity, reliability, cost/feasibility and acceptability of such an assessment tool. The question remains how this will influence or impact assessment in this setting moving forward, and many factors including the availability of resources will influence this.

As mentioned, at present the Rwandan internship has no set 'pass/fail' standards but is rather 'structure-and-process' based in that once the year of internship is complete, independent practitioner status as a General Medical Officer (GMO) is given. The Postgraduate Medical Education WFME Global Standards 2015 suggests a number of basic and quality development standards for the Assessment of Trainees. This includes formulating and implementing a policy of assessment, using a complementary set of assessment methods and formats, and stating the criteria for passing examinations and other types of assessment and evaluating the reliability, validity and fairness of assessment methods (World Federation for Medical Education, 2017).

While these lofty recommendations have been made at an international level, little published literature exists about introducing competency-based training and/or assessment at the postgraduate level in global limited-resource/emerging settings. One study compared competency-based evaluation using Patient Management Problems (PMPs, or clinical vignettes of emergency and common clinical scenarios) with more global, subjective supervisors' evaluations of a large cohort of residents. When globally/subjectively evaluated by their supervisors, the

majority or residents were judged “competent.” Less than 2% of residents were found competent when more standardized, explicit criteria from the PMP scores were used. (Al-Chalabi et al., 1983) The study highlights at least two issues: one, the need for multiple types of assessment in a competency-based program of assessment, likely including standardized, direct-observation based tools. Two, according to the authors, the imperative to move from a subject-and-knowledge centred curriculum to an application of knowledge and skills based training. Both of these concepts point to the direction of CBME.

Our study adds to the limited literature about PGME assessment in limited-resource/emerging settings. Similar to the study by Al-Chalabi et al., this OSCE that embodies many of the CBME principles in its design has uncovered that many graduates that may be passing through the structure-and-process system may in fact not be able to demonstrate competency to the expected level, given that the means of all performance scores were below ‘entrustable for independent practice’. It suggests the need for a robust assessment program with multiple measures and consideration to standard-setting. It may further suggest implications around training to ensure that it is geared towards appropriate knowledge-and-skills application. The OSCE now has a foundation of expertise in terms of local Rwandan personnel who have been involved in a variety of capacities. Should there be a plan to sustain it as a learning assessment tool, workshops and mentoring for local leaders/champions could be facilitated. With the right resources and political will in place, we have demonstrated through the above the OSCE has some sound evidence behind it (in terms of validity, reliability, cost/feasibility and acceptability) to support its integration into a program of assessment in a limited-resource postgraduate setting.

Preface to Chapter 3

In Chapter 2 we evaluated the 'transplantation/export' of the Western-derived OSCE to Rwanda using Van der Vleuten's utility framework of 'VRECA'. We relied largely on quantitative methods in the form of descriptive and psychometric statistics from surveys and scoring data, along with some narrative observations, in order to do this. We presented evidence to support the content and face validity, high reliability (in the form of Cronbach's alpha and item-total correlations), relatively low cost estimates, general feasibility and acceptability of the Western-derived OSCE as an assessment tool for intern physicians' in the culturally different and resource-limited setting of Rwanda.

Approximately one year after the first iteration of the OSCE, a focus group was held with the main objective of exploring in greater depth OSCE examiners' perspectives on the interns' OSCE experience specifically and on intern physician assessment in general. This focus group formed the basis of what is presented in the next chapter. Chapter 3 takes a qualitative analysis approach with primarily non-numerical data to explore the meaning and impact of the OSCE experience. Findings are presented in a visual model and in major/minor themes, then discussed within the 'VRECA' framework to revisit the concepts of validity, acceptability, cost/feasibility and educational impact.

Chapter 3: Exploring the Meaning and Impact of a Standardized Assessment Cross-Culturally – A Trustworthy Physician Assessment Demands Trustworthy Training First

Introduction

Background

Assessment of clinical performance in experiential training programmes relies heavily on judgments made by individual assessors and these may be influenced by socio-cultural factors. The use and adaptation of instruments from one context to another must be considered with respect to cultural validity and specificity in order to optimize the utility of such instruments. Standardising assessment methods and improving assessor training could potentially minimize the subjective bias that arises in clinical assessment. However, there is a significant gap in knowledge of the cross-cultural use, validity and perceptions of assessing clinical performance outside Western countries. (Patel & Agius, 2017)

The present study aims to bridge this identified gap in knowledge. A clinical performance assessment tool (the OSCE), popularly used in Western countries, was 'transplanted and implemented' in the new setting of graduating intern physicians in Rwanda. Following this, the presently reported qualitative study was undertaken. A variety of data sources were used to explore the acceptability (defined as the entire belief system of people in relation to assessment or an assessment method), validity and perceptions of this method and more generally of intern physician assessment, in this novel setting.

Context

In the Rwandan medical training system, after completing 6 years (recently shortened to 5 years) of medical school, graduated medical doctors enter a one-year internship. During this time they are deployed to District Hospitals across the country where they serve in all departments of

the hospital. It is their final required phase of training before receiving a license for independent practice. It should be noted that currently, there is no defined assessment or evaluation standard that qualifies a candidate to move from internship to independent practitioner as a general medical officer (GMO); it is largely a time-and-process based system where upon completion of the internship year, independent licensure is automatically granted. A logbook of clinical exposure is to be kept during internship but is variably used or reviewed for evaluative purposes.

In 2016, Western University in London, Ontario, Canada entered a five-year partnership project with the University of Rwanda under the Maternal, Newborn, Child Health (MNCH) program sponsored by Global Affairs Canada (GAC). The overall aim of this project is primarily to enable training and capacity-building that would strengthen MNCH care in Rwanda and thereby reduce maternal and child morbidity and mortality. The project is called Training, Support, Access Model (TSAM) in Maternal, Newborn and Child Health (MNCH) in Rwanda. At that time, the Rwanda Medical and Dental Council (RMDC) approached TSAM with concerns about an alarming and rising number of medico-legal cases that involved intern physicians and newly certified doctors. The majority of these involved perinatal care. The impression of the RMDC was that there were significant gaps in the internship program and they requested TSAM to provide some assistance in strengthening it. TSAM sponsored a stakeholders' meeting in which an internship strengthening plan was presented by Rwandans that would be based upon formulating and implementing a) internship site selection criteria; b) mentorship of interns; c) interns' refresher course in MNCH; and d) monitoring and evaluation.

As part of principle d), in conjunction with RMDC who administers the internship program, we designed and implemented the first ever Rwandan Interns' Objective Structured Clinical Examination (OSCE) in 2017. It was held in 3 of 5 provinces in Rwanda, had 10 stations per exam

administration, 10 examiners at each of 3 examination sites with a total of 55 interns that participated. During the planning and implementation of the OSCE, extensive consultation was undertaken with Rwandan partners in developing the exam. Robust training was also conducted for examiners, facilitators and standardized patients. The details of this process, along with performance score results and psychometric analyses, are reported in Chapter 2. Of these details, the particulars of the scoring system bears mention here as it is relevant to the analysis. For each scored item on the OSCE, the same global rating scale was used. The 6-point scale is based on 'entrustability' and was defined as follows:

A (1)	B (2)	C (3)	D (4)	E (5)	F (6)
NOT ENTRUSTABLE For independent practice	BORDERLINE NOT ENTRUSTABLE For independent practice	APPROACHES ENTRUSTABLE For independent practice	ENTRUSTABLE For independent practice	PROFICIENT For independent practice	EXPERT For independent practice

Entrustability was defined as 'readiness to safely perform the activity without supervision' (Englander & Carraccio, 2014; O. Ten Cate, 2016). The OSCE examination was timed as an exit-exercise at the end of the intern training year as the candidates were about the embark on independent practice. It would therefore be reasonable to expect candidates to be performing at a level that is entrustable for independent practice (i.e., score of 4), although an a priori 'pass' standard was not set for this exercise.

Methods

The COREQ consolidated criteria checklist structure for reporting qualitative studies will be followed here (Tong, Sainsbury, & Craig, 2007).

Domain 1: research team and reflexivity

The research team conducting the study was comprised of Dr. Amita Misir BArtsSc, MD, MSc candidate (Principal Investigator, PI and referred to as AM going forward) and Dr. Sandra Monteiro BSc, MSc, PhD (co-investigator and research supervisor, referred to as SM going forward). AM primarily conducted the focus groups and informal debriefing sessions that contributed to the data (see 'study design' next for details on data sources). At the time of the study, AM was a Canadian-trained and certified female physician practicing as a specialist pediatric emergency medicine physician at an academic medical centre in London, Ontario, Canada. She also held an academic appointment as an assistant professor as a clinician-teacher at Western University, Schulich School of Medicine in London, Ontario, Canada and was a candidate in the Health Research Methodology (HRM) program at McMaster University, Hamilton, Ontario, Canada. This research was undertaken as part of her HRM degree thesis work. Her coursework included a full-term course on Qualitative Data Analysis & Interpretation. She had also undertaken a qualitative research study many years prior as a medical student, results of which had been presented in peer-reviewed conferences in posters and presentations. During the course of the data analysis, AM also sought consultation with qualitative data analysis expert Dr. Sandra Moll M.Sc. (OT), PhD for guidance on methodology.

The focus group and informal debriefing sessions (main sources of data) were held with Rwandan physician examiners. AM had established relationships with many of these individuals before the study, in a variety of capacities. As mentioned above, this study was held in the larger

context of the TSAM MNCH project. Through this project, AM had worked with a number of the examiners in any or all of the following capacities: holding professional development workshops related to health professional mentorship or examiner training, collaborating with them during OSCE development, planning with them to develop local mentorship and training programs for Rwandan health professionals and working alongside them during the OSCE exams. All participants knew AM's professional background and knew of her as a committed member of the TSAM project and objectives. They were aware that this particular research study was being undertaken with the aim to help evaluate and improve the Rwandan internship program.

With AM being a Canadian-trained physician in the context of a federally funded project operating in Rwanda, it is important to note potential biases and assumptions that could influence both data collection and analysis. AM's position as someone that is part of a project that is bringing major funding to the Rwandans may have inhibited their willingness/ability to voice critical views. In addition, coming from a distinctly North American tradition of physician training and assessment, this may bias data collection and analysis towards constructs that support familiar perspectives and traditions e.g., performance and assessment driven training and rigorous national certification processes based on the same.

Domain 2: study design

The chosen approach to collecting and analyzing the data was that of interpretive description (ID) as described by Thorne et al. (Thorne, Kirkham, & O'Flynn-Magee, 2004). ID '…acknowledges the constructed and contextual nature of human experience that at the same time allows for shared realities' (Thorne et al., 2004). It '…provides direction in the creation of an interpretive account that is generated on the basis of informed questioning….which will ultimately guide and inform disciplinary thought in some manner.' (Thorne et al., 2004). Other

features of ID include a presumption that there is some theoretical knowledge or clinical pattern observation that will likely set the preliminary 'analytic framework', although the inductive process that follows should not be confined by this preliminary framework. Lastly, Thorne highlights that in ID, there is a "pragmatic obligation" to assume that the researcher's findings might be applied in practice and that they should therefore be accessible and useful to practitioners in the discipline (Thorne et al., 2004). These features of the ID approach were well-suited with the context and purpose of the present study. There is much literature and personal experience regarding assessment in postgraduate medical training and some of this certainly informed the data collection and analysis *a priori*. In particular, AM's past postgraduate medical education both as a trainee and more recently as a supervisor has been characterized by many tenets of CBME: close supervision with progressive independence, structured teaching and assessment with significant emphasis on knowledge application and based on direct observation, clear learning objectives and defined standards in summative evaluation for progression and certification. SM has also had significant exposure to concepts in CBME. The concept of trust/entrustability in physician assessment, which is closely related to CBME, also influenced the data collection and analysis process. Lastly, our pre-existing theoretical evaluative framework of assessment utility and the associated elements of validity, reliability, educational impact, cost/feasibility and acceptability were also brought to the analysis.

These two concepts (CBME and the utility assessment with 'VRECA' framework) almost certainly informed our epistemic state in data collection and analysis. It is likely that we would have some tendency to look for and validate both theories – that the tenets of CBME are true and good (otherwise, what would that say of the medical education experience of AM?) and that the 'VRECA' model for assessment of utility was useful and adequate when evaluating assessment

cross-culturally. Yet, as per the philosophy of ID it was important to maintain an inductive approach that would allow the 'constructed truths' to be derived in this particular new and foreign context. Also per ID, given the practical context of the research study, it was important to keep in the forefront that the findings synthesized will almost certainly be presented back to Rwandan stakeholders including the Rwanda Medical and Dental Council (RMDC) and Ministry of Health (MOH) who would be looking for practical information to guide their decisions on how to proceed with assessment in internship.

The original research question guiding the study and analysis was, 'What is the experience of and perspectives on postgraduate physician assessment for physician examiners in a limited-resource setting?' The method to answer the question was originally proposed as an analysis of the data gathered from a semi-structured focus group held with physician examiners that participated in the interns' OSCE. As data was produced and analysed, an emergent design took shape. Emergent design refers to the ability to adapt to new ideas, concepts or findings that arise while conducting qualitative research (Pailthorpe, 2017). When it became apparent that the participants in the focus group spent as much or more time talking about training as they did about assessment, the scope of what was to be reported became more expansive to encompass many themes about the internship training experience itself. To strengthen methodological rigour and triangulate emerging themes using additional data sources, a few additional relevant data sources were incorporated, including examiners' informal verbal debriefings with field notes and their narrative comments from paper surveys that had been administered immediately following the OSCE exams, as well as an electronically administered former interns' survey about their internship experience. The participant selection, setting and data collection of each of these sources is summarized below in Tables 10-12.

Table 10: Examiners' Focus Group (primary data source) - details of participant selection, setting, data collection

Data Source – Examiners' Focus Group	
Participant Selection	▪ Convenience sample, all examiners (n=30) from 2017 OSCE iteration invited to participate via email invitation +/- face-to-face reminder ▪ All those that responded that they were able and interested to attend were accepted to do so, with a total of 10 examiner participants with no refusals or drop-outs
Setting	▪ Focus group held in English, in the meeting room of TSAM offices in Kigali, Rwanda (capital city) on May 20th, 2018 (approximately 9 months after first OSCE exam) ▪ All examiners and focus group participants were specialist physicians with 7 men and 3 women ▪ Kinyarwanda was likely first language of all participants, although all also had a good working command of English ▪ 2 Canadian physician moderators (AM & CK) and 1 TSAM manager (AU) were also present with minimal input ▪ 1 additional Rwandan research assistant (GM) also present
Data collection	▪ First, descriptive results from the OSCE scores + post-OSCE examiner paper survey data was presented to the examiners ▪ Secondly, a semi-structured format was used with 7 questions (not pilot tested) designed to prompt discussion around the OSCE results and experience, as well as more broadly around intern physician assessment in Rwanda (see Appendix A for interview guide) ▪ Entire focus group discussion (i.e., after the results were presented) was audio recorded and lasted approximately 90 minutes, field notes were also taken by CK and GM during the focus group ▪ Audio recordings were transcribed by a Canadian commercial transcription service ▪ AM, who was also the primary moderator of the focus group, listened to the audio recordings repeatedly while proof-reading the transcripts and made extensive edits/revisions to fill in gaps where the commercial transcriber had been unable to identify what had been said ▪ Limitations of the original commercial transcription were likely due to unfamiliarity with the content matter, local accent and fluency of the participants and at times the sound quality ▪ Transcripts were not returned to participants for comment and/or correction before analysis

Table 11: Examiners Post-OSCE informal debriefings + paper surveys (additional data source) – details of participant selection, setting, data collection

Data Source – Examiners Post-OSCE informal debriefings + paper surveys	
Participant selection	▪ All examiners from both iterations of the OSCE (August 2017 & August 2018, total of 6 sites) were invited face-to-face to participate in both informal verbal debriefing and paper survey immediately following the exam they had just participated in ▪ n=60 responses total (note: many of these may have acted as an examiner at more than one site and/or more than one year, and so would be 'double-counted') ▪ None refused to participate or dropped out
Setting	▪ Data was collected at the site of the exam (held at a hospital), usually in a meeting room, immediately following the OSCE exams in August 2017 and August 2018 ▪ The PI (AM) and a simulation specialist (KJ) heavily involved with the OSCE were present, as well as usually a Rwandan research assistant (GM) ▪ All examiners were specialist physicians and included both men and women
Data collection	▪ On arrival to the meeting room, a brief paper survey was provided to each examiner for completion (see Appendix B); only narrative comments from this survey were utilized for analysis in this chapter since controlled-response option questions have been reported elsewhere ▪ This was followed by an informal verbal debriefing, where each examiner was asked in turn to make comments about their station as well as any general comments that they had about the OSCE ▪ Written field notes summarizing responses during the verbal debriefing time were taken, usually by KJ or GM or occasionally AM (audio/video recordings not done) ▪ The paper survey + informal verbal debriefing generally lasted about 60-75 minutes at each site, with 10 examiners at each site and a total of 6 sites ▪ Notes and computer-entered data from surveys were not returned for comments and/or correction before analysis

Table 12: Former Interns' Post-Internship electronic survey (additional data source) – details of participant selection, setting, data collection

Data Source – Former Interns' Post-Internship electronic survey	
Participant selection	▪ Convenience sample, all interns who had participated in the OSCE in 2017 and 2018 (n=104) were invited via email to complete an electronic survey about their recent internship experience (see Appendix C) ▪ Note that since the OSCE was run in 3/5 provinces, not all interns in each year participated in the OSCE; usually about 60% of the entire intern cohort for a given year participated in the OSCE ▪ A total of n=61 respondents completed the survey ▪ None actively refused or dropped out, although about 40% of those invited did not complete the survey
Setting	▪ The survey was sent to the entire sample at the same time ▪ Respondents could complete the survey at their choice of location and time during a 6 week period ▪ Data was collected between October 2019 – February 2020, which would have been approximately 1-2.5 years after completion of their internship
Data collection	▪ The survey was structured into sections and consisted mostly of controlled responses from available options/likert scales ▪ At the end of the survey, an open-ended narrative comments box was provided where respondents were invited to share any additional comments about their internship experience ▪ The survey was in English and reviewed for local content validation and relevance by Rwandan physicians and interns before distribution ▪ The PI (AM) sent the survey (and reminders) via email distribution using Qualtrics survey software through her home institution (Western University, Canada)

Domain 3: analysis and findings

For the analysis phase, two coders (AM and SM) were employed. After initial reading of the data, a common codebook with definitions of all codes was formulated and adapted before coming to a final version. Primarily topical and analytical coding was used, in addition to use of memos for emerging ideas and annotations for key quotations. (Richards, 2015) Themes were derived from the data and codes and discussed to reach consensus on interpretation and

representation as they are presented in the results and discussion sections below. As previously noted and in keeping with the ID approach, interpretation and representation was influenced by pre-existing theories of CBME, the concept of trust/entrustability in physician assessment and the VRECA assessment utility framework, although our results and conclusions were not bound by the original tenets of these concepts. One example of this is the concept of entrustability, which is typically thought of in the context of assessing and trusting the competence of an individual physician. In our instance, the importance of trust was found to be relevant not only in an assessment of an individual physician, but more expansively in the training and assessment process that forms the basis of trust in any individual physician. NVIVO 12 Plus (version 12.5.0.815) software was used to manage data and coding. At the time of writing, member-checking of the final analysis had not been undertaken.

Using the above outlined emergent design, additional data sources and analysis methods, data saturation as it related to the original research question was approached. The number of new ideas or themes emerging diminished, existing ideas/themes were strengthened by triangulation of data sources and 'negative cases' (i.e., data appearing to contradict apparent trends/themes) could be reconciled within the theoretical interpretation/model. Having said that, a few methodological limitations warrant mention:

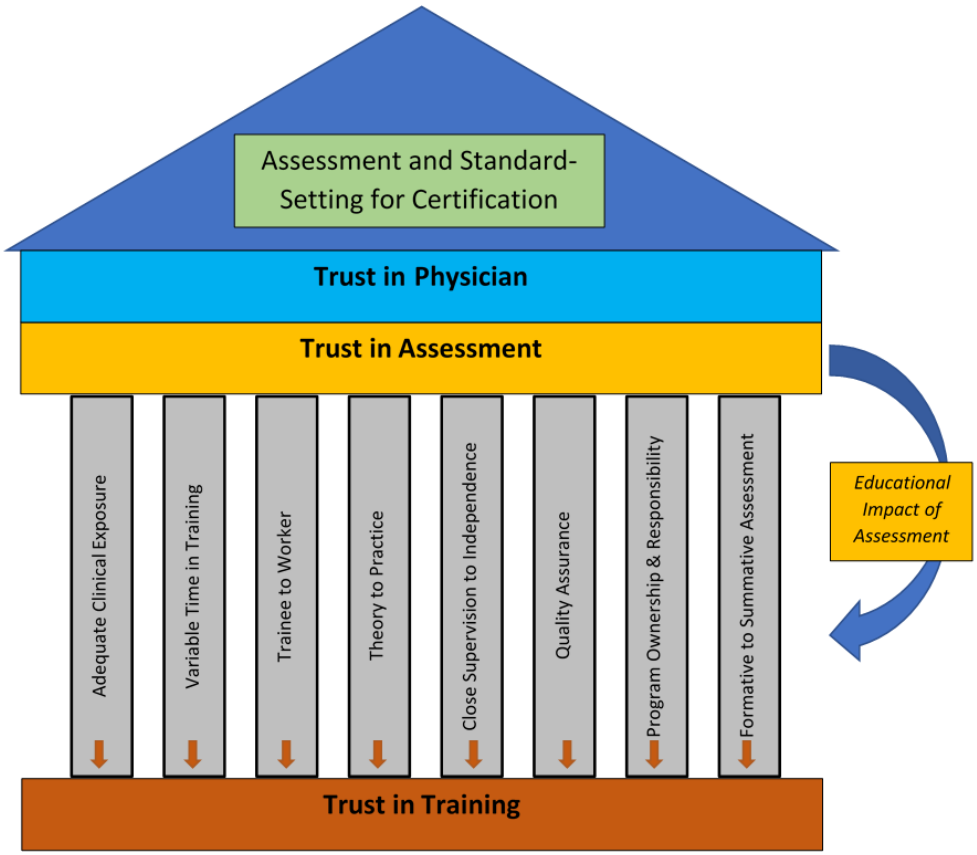
1. Participants were selected as a convenience sample in all cases. Purposive or theoretical sampling was not undertaken.
2. The focus group included only OSCE examiners. Other relevant perspectives (e.g., interns, medical staff from internship site hospitals) had limited (i.e., through highly structured electronic surveys) or no opportunity for more in depth exploration.

3. Field notes were made but limited in their detail. Who was talking at each turn was not always identified. Therefore the quotations provided in the results section do not specify which individual was speaking.

Results

After analysis and integration of the various data sources as described in the methods above, a descriptive model was generated to summarize the findings. Figure 4 displays this model. The central theme that emerged was that an assessment may be valid, reliable, feasible and acceptable, as demonstrated in the previous chapter. However, in order for its results to be trusted as a true reflection of physician ability and therefore used as a standard-setting process, physician training must first be seen as trustworthy – which in this case, it was not. This is depicted visually in Figure 4 as a house: a trustworthy training process serves as the foundation (brown) and the features of a trustworthy training process are the pillars (grey) that contribute to that foundation (indicated by arrows pointing into the foundation). Only if this is established can an assessment be trusted as a true reflection of the physician's ability (yellow), which in turn supports or refutes the entrustability of an individual physician (light blue). The exterior roof of the house – the final, important layer of protection before exposure to the patient environment – is utilizing assessment and standard-setting for certification.

Figure 4: The House of Trust – Building a Trustworthy Physician



Building Materials:

Time
People
Purpose
Expertise/Training for Supervisors/Examiners
Money

Two important observations are worth noting with this model. First, a roof without a house is useless, and a house without a roof is equally useless. Therefore, addressing or implementing assessment can not be done without also addressing and implementing proper training. Similarly,

implementing proper training can only be considered complete if some form of assessment and standard-setting for certification is also undertaken. Second, all parts of the house – the foundation, pillars, subroof, external roof – must fit together. It is best to think of this 'fit' as common purpose: the training should be geared to the assessment and certification and vice-versa. All of which should presumably be geared to the physician's ultimate scope-of-practice.

Two ancillary components to the house are included in the figure. One is an arrow drawn from 'Trust in Assessment' to the pillars of training, with the label 'educational impact of assessment'. This represents one of the important reflexive effects of an assessment that was identified by study participants, that is to identify deficits in training (at both an individual, but more importantly, programmatic level), address them and thereby strengthen the training. The other component outlines the 'Building Materials' for the house. As discussed by participants, a number of 'raw materials' are critical in the success of training and assessment: time, people, purpose, expertise/training for supervisors/examiners and money. Although the challenge of acquiring most of these 'raw materials' in sufficient amounts is not restricted to Rwanda, it is certainly exacerbated there and likely in other limited-resource settings.

The remainder of the results section will go through specific proposed components/relationships of the model in Figure 4, describing in greater detail what it means, how it came to be incorporated into the model, and where appropriate, providing supporting raw data. It is worth bearing in mind that the 'raw data' was often conversation in the context of a different culture with its own style of communication, with English as a second, possibly third language for most. The authors have taken care to maintain the integrity and authenticity of the participants' discussion. Therefore we have avoided paraphrasing or revising participant comments when

provided directly from transcripts, prioritizing this over the brevity often sought in Western communication styles.

a. Training, assessment and certification are integrally linked

While the discussion in the Examiner's focus group began with the questions about assessment results, process and standards, the comments quickly and continually returned to the quality of training in addition to the importance of assessment and certification. Results of below-entrustable mean scores on every OSCE station were presented to the focus group at the start of the session, and the first question posed to them was what they thought of the results and whether they trusted them as an accurate reflection of clinical performance for the interns. Rather than either doubt the content validity of the OSCE results (i.e., believe that the test did not accurately measure the underlying construct of a competent independent practitioner) or conclude that most of the physician trainees were not trustworthy of independent practice, virtually every individual pointed out or concurred that this reflects a problem in the training. The following quotation from one of the examiners highlights the progressive and dependent relationship of training (both before and during internship), supervision, assessment and certification:

…the first step is actually also to be part of knowing actually if the schools that are training doctors actually are following actually the kind of curriculum that is actually being monitored and see if actually it's making them practice with some courses to be covered, if they are really really covered and so on. And then, later, also having a kind of permanent supervision also, even at the hospitals that are hosting those training sites to see if they're really fully equipped, if they have actually needed materials actually to host those junior and senior track training program students, and then the same also for the internship program sites. And then, following them and for sure using also this kind of OSCEs also as one part of actually assessing because you need actually to be tough because someone who is going to be treating patient's lives, people need actually to be really competent, and this is also one of the ways to know who needs the knowledge, if they have the knowledge too…'

Another examiner discussed the importance of ownership of not only assessment and certification, but also of supervision and training. The suggestion was that the RMDC must be more invested

and involved in both supervision/training and assessment/certification. They can not implement one without the other:

'I think for the Rwanda Medical and Dental Council may be able to use OSCEs as a part of licensing, it has to assume some other responsibilities like to be fully involved in supervision and training of interns. So there's no way you can evaluate some people that have been one year somewhere, and you don't know them, you don't know what they are doing, and then they come to the end of the year, for assessment and evaluation. So, if Rwanda Medical and Dental Council can fully assume responsibilities of adequate training in the, like, in standard ways, then it can take over to use the standards, the evaluation form for interns to pass for licensing...'

b. Training and assessment is a process over time

The need for training and assessment to be seen as a gradual process during internship was represented as a variety of different transitions: theory to practice, trainee to worker, close supervision to independence and formative to summative assessment.

Theory to Practice

Many comments from the examiners talked about how interns have knowledge but are lacking in application in practical skills. This was perhaps most obvious for technical skills e.g., suturing, neonatal resuscitation, balloon tamponade for postpartum hemorrhage, however it also applied to other clinical scenarios. One examiner captures the idea:

'I think for this question, [on being] entrustable (background noise) I think it maybe, it must be emphasized on it. Because when we see that result, we see for that OSCE examination, see that the point we see on the result, it means the…the practice is very low for the interns. It meant the interns, they know more theory than practice. And we see internship, it has to emphasize on practice because the intern must be independent in practice.'

Comments from graduated intern physicians also mentioned this transition and suggested that at times they felt it was done well, at other times could be done better. A couple of interns wrote how *'The interns should be more trained on practical skills in gyne-obs, minor surgery and emergency care'* and how *'Internship helps to relate the theories and practices'*.

Trainee to Worker

It was proposed that the interns need to first be seen as trainees/learners, and then gradually be given responsibility as unsupervised workers. There was a feeling that currently, many institutions do not appreciate this as stated by one examiner:

'So, it is site-dependent, and we still have a big way to go, especially struggling on who is an intern, being understood from the Ministry level up to the Hospital Director's level. Some hospitals still look at interns as doctors, who are coming to work, they give them stamp – go…So, I thought intern and this internship is looked at as a stage, a transitional stage, of a student crossing over. And if it cannot be seen in that context, it will remain controversial. Because people want interns as workers. Not people who are coming to learn.'

This lack of clarity regarding role is supported by former intern survey results, where 47% of graduated interns indicated that they did not feel the hospital(s) where they worked were clearly aware of their role and limitations as an intern. Additionally, 38% felt that the level of responsibility expected of them was not always appropriate for their level of experience/expertise during their internship.

Despite this tension between trainee-versus worker, there is a recognition that there is a need to strike a balance between the two as evidenced by this exchange amongst the examiners focus group participants:

'You can learn by working.' <Laughter from group>
'Your certainly can.'
'Because they did much of their learning, they are also working. And they really help hospitals much, and so, it should be in the interest of the hospital to host interns. And by hosting interns, they should be positioned in a way that will help them and also interns will want to go there.'

Close supervision to Independence

One of the underlying reasons for inadequate performance on an assessment may be inadequate supervision during training. Below were the comments of examiners in the focus group:

'Because if you have some interns who are not able to manage something which is common in district, it means that the supervision is also poor. I think in both side [intern performance and supervision], we have also to do something to change the system, to change the supervision somehow…'

'For those that have been in internship sites, interns are not supervised, still they are taken as workers who are given stamps to go and start practicing without any supervision at all.'

'…do you think [supervisors] get enough first-hand exposure to interns actual practice? Actually, the way it is now, it's very difficult, because…I don't know the statistics but maybe 60% of the interns, they work like as independent practice in district. So…they are supervised somehow, but you may seem like consultant in the office of consultation, and that office is like a full general practitioner. So, it's not like direct supervision, it's like an indirect supervision but it's somehow the exposure for the supervisor and their interns because our GPs are still few…'

Additional comments from the informal examiner debriefings that immediately followed the OSCE exams and in the post-OSCE Examiners' Questionnaire similarly talked about the need for better supervision, more active teaching support and need for constant mentorship of interns.

Survey responses and comments from the former interns also suggested that there was room for improvement in supervision and teaching, although it did appear that most felt there was some basic level of support. Around 75-85% indicated that in situations where they felt they needed to call for help, there was always someone to call (day or night) and that help was readily given in a supportive way. 60% of respondents agreed to some degree that the more senior physicians took initiative, time and interest to teach and guide them, and that their clinical supervisors were often present and readily available for case review and/or direct supervision. About the same proportion reported that at the start of their internship, they were assigned to the night shift with a more senior physician so they could learn from them and ensure safe patient management. In the open-ended comments box on the survey, several graduated interns suggested that supervision needs to be more systematic, regular and supportive, including one comment as follows:

'During my internship periods, the issue of supervision was not applicable as we had no seniors to follow our daily progress and help us to improve on our weakness. We tried ourselves to put in practice the theories we got from medical school and we ended it well.'

Like the trainee versus worker balance, the issue was raised of what is the right balance of supervision versus independence, and whether interns would get better experience at a designated teaching hospital versus district hospitals. One examiner raised the issue as follows:

'It's a big thing actually, the interns don't want to be in teaching hospitals because they say it is like they track [specialty] residents, they are scrutinized in Rwanda, they say it is like they are doing their final at (inaudible) medical school. And they prefer to go in a place that is unsupervised and try…"discovering" <hand gestures, group laughter>…So, I thought this coming meeting on internship, it will be very interesting. Because some people (inaudible) think, say no no no no, no interns at teaching hospital. They are competing for patients. Is it good to see 100 patients and mismanage them or to get twenty and to treat them properly?'

It is also notable that a few comments from graduated interns mentioned how they 'ended it well', despite no or minimal supervision. This would suggest that there may be some value to allowing trainees to learn independently, through self-reflective experience and increased clinical exposure versus close observation and feedback. This balance must be carefully considered from a patient safety perspective however, as approximately 30% of former intern survey respondents indicated that there were times that they felt that a patient may have been put at increased risk of disability or death due to their relative lack of experience and/or lack of appropriate supervision.

Formative to Summative Assessment

The topic of assessment was central in the focus group discussion. A broad range of comments arose when addressing this complex area. There was a recognition of the need for formative assessment leading up to summative assessment, with some recognition that this does not preclude the use of formative assessment data being used at a later point to make summative assessment decisions, a process and tension that has been recognized in the literature (C. P. van

der Vleuten et al., 2012) Several principles became clear when reviewing assessment in the context of internship. This included:

- Need for formative assessment and feedback at regular intervals during training in addition to exit or summative assessment
- Administering the OSCE to be as free from biasing factors (e.g., language differences, SP role portrayal, clarity of instructions/prompts) as possible
- Inclusion of multiple sources of data during the assessment process to get a complete picture of performance
- Establishment of a 'heterogenous team' of assessors to limit bias
- Utilization of the concept of entrustment, based on a collaboratively built understanding of what this means, to underpin physician evaluations

In the focus group, debriefings and survey, the examiners noted that the OSCE exercise was quite new to most interns, thereby possibly influencing their performance in a negative manner simply due to unfamiliarity of the exam, particularly with use of relatively novel simulation-based methods such as standardized patients and mannequins. There was also comments about the fact that interns were not told that they would have such an exam at the end of their training and how that could also result in under-performance due to lack of preparation. Some commented that many interns may have French as their primary functional language and so perhaps the OSCE should be translated into French. Some indicated that the information given by the SPs was sometimes inconsistent. Several noted that the practice of calling/asking for help from a senior or specialist physician was not common cultural/professional practice in Rwanda and therefore commonly missed on the exam. Finally, it was noted by some that the fact that this was 'just an exercise' and not a true summative evaluation with real consequences, they got the impression that

at least some candidates did not take the exam seriously. There was some uncertainty and controversy over just how much these factors contribute to the performance of the candidates or bias of the scores, if at all.

To help address some of these potential sources of bias, as well as to utilize the OSCE as a learning/feedback and not just an assessment/evaluative tool, the suggestion was that an OSCE should be done at least once at the mid-point, more preferably every 2 weeks to 3 months, during the internship and even medical school training years before using it in any summative evaluative manner or decision. They also talked about the need for 'continuous evaluation' of the intern locally throughout internship, where that may be based on a combination of a logbook, standardized quarterly evaluation forms, locally-based OSCEs or other assessment tools. This 'continuous evaluation' data as well as a summative/exit OSCE scores would each contribute a certain proportion to inform a summative evaluation decision at the end of internship. There was suggestion that OSCE scores, in addition to continuous evaluation, should be used to generate a recommendation as to where intern physicians will be appointed and that those recommendations should be taken into consideration before they are given full responsibility to practice.

One examiner was particularly thoughtful about the risk of bias and inaccuracy if all assessment was left to the local supervisor(s). He comments:

'So, I think, to me I think it would be better with a team, maybe a heterogenous team [for assessment]. So, someone who has been a supervisor for one year for the intern is good for continuous assessment. But sometimes there is emotional and a few other things, so that can say…maybe this intern is good, ah? (laughter from group) So, to avoid that, I think maybe if it's a team of three people composed of like a continuous supervisor who has been with interns, and some external examiners like one or two, then it can make a heterogenous group to make a maybe a suitable decision for the intern. Otherwise…if the ball is for the supervisor, then he will count the daily activities and so and so, and everybody is a human being, he may just…there may be some emotional and some other social bias to judge your intern if you have been with the intern for one year. But if there are say two or three people, some are external and others are just daily supervisors of the intern, then they can make…a middle decision.'

As the OSCE was the first time that the concept of 'entrustability' had been introduced or used, examiners were asked what they thought about it. In general, there was agreement that it was a useful and practicable concept, although some seemed to define it more subjectively while others define it more objectively. One examiner describes his intuitive perspective and the importance of direct observation in determining entrustability:

'So the entrustability aahh, scale, I think it's one of the best way to evaluate trainees in medical career in general. Ah, where when I mark you as entrustable, it means maybe if I'm sick and I see you come with a treatment, maybe I'm satisfied…and when you are not entrustable at all, when I mark you, if I am an examiner and I mark you not entrustable to me is if there are two doors and I'm sick, I can't beat on your door, I will choose another one <chuckles from others>. So, I think [it] is one of the best method we have in place to evaluate trainees in the medical area. And I think it's appropriate, as far as the trainee is performing, showing up his skills or her skills and I'm standing there to see if the way he is performing it on a patient, I can myself, or I can set my (inaudible) to be a patient for that doctor or that trainee, then I'm really satisfied for that. So, if I…if I mark you as entrustable, to me, it means your skills that, as I observed you, is enough for me to go forward for the independent practice. So I think there is no doubt that this is the best way for me to evaluate interns or other people in the medical field.

Another examiner describes a more quantitative understanding of entrustability:

'Also, I found this system is good because it is not saying yes or not. To be in one criteria or in one grade, it has a some requirement. There's no difference saying that she got 70 percent, she got 60 percent. Even if we go in percentages, we already said, to pass it requires 50…I don't find any…difference between the [entrustability] grading or to percentage. Because you have criteria that are set saying if on this question, you respond from A to C, or you respond like this, you are in this grade, it seems that it is like that on this question you got two over five, you got three over five. And at end, they do the sum and you got…I found there's no difficulty in saying this maybe is good for grading. Even, because...if it is a grading, it means if you say it is entrustable, it means…you are not scored at 100%, nor you are not at 20%. You are medium to go and to help people. I find it has a good meaning that we can adopt it.'

In both cases, whether arrived at through a more mechanical or intuitive process, the end result is a judgment indicating whether or not practice for the observed task is sufficient to manage patients independently. In addition to agreement on this fundamental meaning and applicability of entrustment, the examiners emphasized the importance of experts pre-discussing standards of entrustability to come to a common understanding about it:

'It was not difficult [to decide if a candidate was entrustable or not] because the criteria to follow. So, if you have some criteria set for each medical condition and if the intern did not fill the criteria set so he or she was not judged to be entrusted. So, the criteria are there and clear, it will be easy for every examiner.

(Moderator) You mentioned the importance of experts sort of discussing and getting common consensus. Did you find that was helpful in identifying entrustability?

Yes, it's helpful. Because for each specialty has its way of evaluation. For example, for us, if a trainee can come, examining patient with a….maybe PPH [post partum hemorrhage], and he forgot to call for help, the patient can't be managed with one person only. So, calling for help would be crucial for management of that patient. So, each specialty has its (inaudible) examiner setting the criteria.'

The post-internship survey administered to former interns included some questions around feedback and assessment. The results were mixed around the current practices of timely, useful evaluations at regular intervals throughout internship and supportive of the OSCE being a useful learning experience and assessment, as summarized below in Table 13. Available narrative comments were in keeping with the survey statistics and examiner recommendations, suggesting need for close follow-up, regular assessment and feedback of interns, practice OSCEs and quarterly evaluation at sites.

Table 13: Former interns' survey results – feedback and assessment

Survey item (related to feedback and/or assessment)	Disagree (%)	Neither agree nor disagree (%)	Agree (%)
Received regular, timely on-the-job feedback (written or verbal) on my performance throughout internship year.	35	13	52
Feedback provided in manner that was helpful in identifying strengths and weaknesses.	25	16	59
Supported in trying to improve weaknesses	20	15	65
Evaluations were fair and reflective of my abilities	20	18	62
Interns' Logbook was a useful and practical way of continuous evaluation during internship	33	20	47
OSCE was a useful learning experience	9	5	86
OSCE was a good method of assessment	15	7	78
OSCE exercise should be run for future interns	4	6	90

Two other important principles about the process of training and assessment that were noted in the examiner focus groups, debriefings and post-OSCE survey were the need to ensure adequate clinical exposure for interns and the appreciation that getting to the end-goal of an entrustable physician may require variable time.

Several comments were made about how interns with less exposure to certain procedures or disciplines would do more poorly on the related areas on the OSCE. The suggestion was that ensuring structured rotations in all relevant disciplines, with an emphasis on ensuring exposure to practical skills, needs to be a goal, as it does not seem that it is currently consistently achieved for all interns.

During the focus group, multiple examiners indicated support for the idea that internship, which is currently time-defined as one year duration before passing on to independent practice,

should be viewed as potentially time-variable and defined by performance-based standards. Two examiners note:

'So, if there was a way of assessing someone who's finishing internship and saying, yes, you are ready to go, yes, but not yet ready, maybe do another six months…such kind of evaluation which is focusing on identifying before letting people to go to practice, maybe it would be helpful in our medical practice…'

'Of course, internship is not ah, is not a market – you go in, you come out. It's not a matter of spending a year there, there should be expectations and that they know in internship, there's minimum expectation that a doctor who passes through internship, should be able to do A, B, C, D. And, of course, with cognizance that they are junior GPs.'

This shift in thinking from time-and-process based to competency-based advancement and promotion would likely require some culture change. However it is not a totally foreign concept and extension of training has some precedence at least in the setting of remediation, as highlighted in the comments of the two examiners below:

'…there must be something to say if you can't attain this, you have to repeat a year or a rotation on a certain discipline. And it's not the part of our culture in terms of medical training unfortunately. And, eh, and so is even undergraduate training. So, it's something that is creeping in now, but it's not been … it brings a lot of friction here and there.'

'I remember even for me that sometime, a doctor can go to the hospital…I remember a doctor went to a hospital. Then worked there for two years, then after that, they found that he was doing the mistakes, then they decided that he should do again the internship.'

c. Quality assurance for the training and assessment process, and ultimately for certification, depends on program ownership and responsibility

Many perceived gaps in the training and assessment process were identified above. These include deficiencies in: practical application of knowledge, appropriate supervision, understanding of expectations and limitations of intern as a trainee versus worker, meaningful formative and summative assessment leading to certification standards, adequate clinical exposure and flexible duration of training. Quality assurance, or the processes and procedures that systematically

monitor different aspects of a program or product to ensure that it meets specified requirements, could be considered one way to meet these challenges. In this perspective, monitoring both the training and assessment process/program and the individual trainees/physicians is relevant.

There was recognition from the examiners that the ownership and responsibility of quality assurance is shared at both the local level, in the form of hospitals that are hosting internship programs as well as the national level, in the form of the Rwanda Medical and Dental Council (RMDC) that is administering the internship program. It was felt that the national level should be taking the first steps and leading the way, while acknowledging that local hospitals must also take ownership of their part. Two examiners make comments about the national level:

'…The Rwanda Medical and Dental Council are just to assess our hospitals and see which hospital can actually host the interns, looking also to the doctors they have and if they are really, I can say, able to make such kind of judgment [as supervisors for interns], also evaluating also, interns. So, I think for me, the first step is that this should be a first step to evaluate the hospitals based on what you need from interns to get after the internship to see where they're going, if there is all needed materials actually to run and to get and to achieve the expectations. And then to see among them, the requirements, is to have actually also senior doctors that can actually be able to evaluate them and then to give them actually a tool to use actually for them [to assess]…'

'But again, looking at…if the intern that is at this hospital, is he having the same opportunities of the same that is under that hospital, so that you can actually keep them same evaluation. So, that's where now the Rwanda Medical Council comes in before, they are appointing the interns to see if at least they can get hospital that if an intern is there, can get actually same opportunities of learning and also, get also evaluated by senior people. And then, at the end, then coming with the OSCEs as also to have part of the training…'

In fact, in 2017/2018, RMDC had created and published a set of criteria/standards that would qualify a hospital to be a proposed internship training site. After the initial implementation of these criteria, internship training sites became reduced from over 30 to around 20 sites. This may be seen as a step in the right direction for quality-assurance.

Challenges remain in terms of enforcing those criteria, and it seems that there remains work to be done in terms of national definition/standardization of what you must provide in terms of

teaching, learning, supervision and assessment provisions as a training site. Local hospital understanding and uptake of their mandate as a training site then also must be facilitated. This is explained by one examiner who was integrally involved in creating and implementing the site-selection criteria:

'Like the internship site selection has a criteria, and one of the criteria…the minimum standard is to have a specialist at least in a department to post an intern there. But then, when we looked around, the hospital didn't have a specialists, so we had to bend down the standard and say you know what, a senior credentialled medical officer could be also be mentored to mentor interns. And the criteria is there, it has a maximum and basic minimum, but as it stood now, at least we know what a hospital should have. But what it means is to have those teams at a facility level and to understand what internship is all about. Because interns from the system are still taken as workers.'

'I think, sorry, internship as it is happening now is not universally understood, even by the senior doctors. One being that they never had internship we are talking about today. And so, they need also empowerment to be able to be actually supervisors. So, when there was a sort of structure of internship, there was a plan to have mentors. And these were to be mentored through the whole structure of this, what it means to…to see the expectation of the intern and the supervisor which is not known. It is individual dependent, there is no standard as of now.'

Several other examiners similarly identified possible confusion or lack of awareness about roles and responsibilities at the local hospital level as a barrier to quality internship training:

'I was thinking that even some hospital may receive the interns where there are some doctor who may supervise them, who can teach them, and they're not aware that they will come and they have such responsibility to teach them. So, I think that we should do, feel that have such ownership that it can be done. Thank you.'

'…I remember there are several criteria that were looked at and…one of the things that an internship site should have, it should have training in its mission. So, it's not that you are a very good hospital, if you have no mission of teaching, you can not qualify to be an internship site. So, as he rightly says, teaching must be part of the job profile for whoever whom works in internship sites. Because they've been found to have in their mission statement that they are there to teach, to educate and all that…'

When there is a suggestion of shared ownership, it raises questions – and potentially confusion - about who is responsible for what and what organization(s) are best aligned for which activities. Between the focus group, debriefing comments and post-OSCE questionnaire narrative comments

from examiners, this uncertainty was revealed. Certain comments suggested that ownership for organizing regular practical training and assessment sessions during internship (like structured orientations to neonatal resuscitation, surgical skills sessions, OSCEs, etc.) should rest with regional referral hospitals and/or district hospitals that are internship sites. Others suggested this should be mandatory for the RMDC to execute. By contrast, others felt that RMDC is not an academic institution, so should not be giving exams, but rather that the focus of training and exams should shift to the medical school and training of medical students.

Despite some apparent confusion around responsibility, there did seem to be a consensus that a uniformly defined, performance-based standard should be set nationally to ensure standards and safe practice for all physicians, including those that may be coming from abroad. This standard should be worked towards gradually and implemented only after fundamental changes in internship training have taken place, as suggested by examiner comments:

'I think a foot has been put down and say, this is the standard, and we work towards that. I see it percolating at different levels, being taken up by Ministry of Health, deciding on who is allowed to have an independent practice. In other places, you have to be of a certain calibre to be allowed even to open your own independent practice…but we don't have those yet.'

'I think…if somebody is not training you and come at the end to evaluate you it is somehow a little bit unfair. But if [RMDC] can assume responsibility so training you and being with you all the one year of internship program then at the end use the same standardized training method and use its (inaudible) standardized way of evaluating you, then I think that side is fair.'

'The OSCE cases can be used as part of RMDC physician licensing after harmonization/standardization of internship centres'

With respect to the former interns' experience of physical facilities provided at their internship sites, about 55-60% responded that they had acceptable accommodations and consistent access to a clean, private, securable room with bed to rest within 10 minutes of the hospital for

night duty. About 75% responded that their hospital had fairly or very adequate equipment, services and facilities available to carry out their duties.

With respect to internship administration and orientation, the picture was mixed. Of note, around 50% of respondents indicated they did not receive the Medical Interns' Handbook from the Ministry of Health at the start of their internship, which includes their logbook that they are supposed to use throughout internship as a main record of clinical training experience. Also of note was that 44% responded that they either did not know their internship co-ordinator at their hospital and/or did not feel comfortable accessing them. Orientation to training objectives and scope of clinical work during internship was reported as adequate for 29% while 40% felt it could have been done better and 30% reported not done adequately or at all. Orientation to hospital(s) and each department during internship was reported as done consistently for 46%, with 42% indicating it was done inconsistently and 12% reporting not done adequately or at all.

Narrative comments from the former interns' survey suggested that better structure, implementation and monitoring would be helpful. Recommendations included that there should be an academically oriented senior supervisor with clear internship objectives at each site, who is responsible for interns. Regular follow-up visits of medical interns at their sites to share experiences and challenges to be solved was also suggested.

d. Quality has requirements and costs

The 'House of Trust' in Figure 4 shows the 'Building Materials' of Time, People, Purpose, Expertise/Training and Money. These were identified, in a variety of ways, by the examiners and graduated interns as necessary for the provision of a quality internship program. Each of these requirements, with details about the nature and challenges of each of them, is summarized in Table 14.

Table 14: Requirements for Building the 'House of Trust'

Requirement	Comments	Challenges
Time	• For close supervision and feedback/coaching • For preparing and executing training sessions • For preparing and executing assessment • For extension of training/remediation where needed	• High student-teacher ratios in medical schools already • High volume and acuity patient demands in clinical settings • Time may be further taken away from patient care • Intern doctors extending training could mean temporary medical officer shortages at district hospitals
People **(for consultation, teaching, mentoring, supervision, assessment)**	• Program leaders and administrative personnel (national) • Specialists (on-site) • Senior medical officers (on-site) • Junior medical officers (on-site) • Internship Director (on-site) • Administrative personnel (on-site)	• Limited dedicated internship program leadership/personnel at national level • Limited specialists at many sites • High turnover of medical staff • No administrative personnel to support program locally
Purpose	• Clarity of and commitment to teaching/training mission • Clarity of and commitment to performance-based certification standards	• Local hospitals and on-site physicians may have little experience in teaching mission • Certification standard setting difficult without quality-assured training and assessment first • High-stakes standard-setting requires rigorous process for decision-making, and needs to be accompanied by options for appeal and remediation
Expertise and Training **(for patient care and teaching missions)**	• Expert general clinical skills for routine care • Expert specialist clinical skills for consultation/management of complicated cases • Teaching, supervision and mentorship expertise • Examiner training • Education leadership/scholarship expertise	• General clinical skills of current medical officers at internship sites may not always be expert or evidence-based/current • Limited specialists to provide guidance around complex patients • Currently no formal training in teaching/mentorship offered • May be limited health professions education experts at national and local levels
Money	• For time and people as above • For supervisor/examiner training • For training and exam materials development and implementation	• No extra available money readily available

Of these requirements, some may compensate for others. For example, if more people are introduced to the system, then less extra time is demanded from each individual. Similarly, if time and/or people are provided without additional cost (e.g., expansion of existing job roles), then less money is required.

It is important to note that while money is often considered the greatest limiting factor particularly in a limited-resource setting, it is not the only nor even necessarily the most critical of the requirements. Several comments from the examiners' focus group highlight this.

'The other thing is that those selected hospitals, that will have the most interns, are they having missions…including teaching, including researches, including – so are the doctors who are working there, I mean this should be in the concept (inaudible) of responsibilities. Like in a teaching hospital, we evaluate a student without requiring money….then I think with those provincial hospitals or referral hospitals, those are doctors who are specialist doctors who will be sent there. This should be among the responsibilities. So, I don't think money or time will be…the obstacle to do this method to evaluate intern doctors.'

'So, as he rightly says, teaching must be part of the job profile for whoever whom works in internship sites…'

Other comments address the requirements of time and purpose:

'Yeah, I want to make a comment about the time. Time of teaching, supervise and complaining of the work. Where there was a will, everything is possible, and it will depend on how the person is organized. Saying that, I can say, if someone is having a will to teach or to supervise intern where he is working, or she's working, I think it can be done. For example, there was the different way you can teach the person, you can teach the intern through the presentation or during bedside teaching. So, sometime when I try to make the analysis myself, it is not a matter of time or high clinical demand work. Sometimes it depends on the person's self-organization as well. So, I think it will take time to have like ownership and think about it as the people who are responsible to do that and I think it can be possible. And probably, we can have even the sharing experience between those hospitals which have been chosen as internship sites so that if a hospital can learn from the other the best practice that they are using so that it can happen.'

My point of view, saying that someone is having a limited resources, I mean limited money, that determines something which will do … will give the standard things for better care of the patient, and it is time-consuming, it doesn't mean that we should not have it in our country. So, whatever it may require, I think it can be done. It can be done for the better of our population. So, for me, maybe I will support that this is such method of assessment can be done in our settings. The way I see that it can be done, the people must really understand that it is really needed, and you should have a kind of the communication, the communication which is clear.

Throughout the focus group session, there was a range of discussion about supervisor and evaluator training as well as continuing professional development. There was acknowledgement that a heterogeneity in supervisors may lead to some heterogeneity in the training experience (e.g., an intern who has a specialist pediatrician as a supervisor will probably not receive the same experience as someone who trains under a senior medical officer or a junior medical officer). However, it was also felt that this did not preclude any one particular category of doctor from being suitable to be a supervisor and/or assessor, so long as those individuals were empowered, trained on internship objectives/standards as well as how to be a mentor and supervisor, and given some standardized tools and processes. There was some discussion also about the quality of clinical skills for those currently in practice at internship sites, with concerns expressed that some even 'senior' people may not be practicing the best, most up-to-date/evidence-based care. There was some suggestion that an OSCE could be used as a periodic practice exercise for those in practice at District Hospitals, to identify and remediate any deficits and to maintain skills and knowledge.

There was also discussion around on the critical issue that if a certification standard was to be introduced and more interns needed further time in training before being certified and 'deployed' as independent practitioners, this may mean less physicians to provide care to patients in an already significantly underserviced settings. The consensus on this last issue however was fairly clear – this cost is worth the benefit. One examiner explains this elegantly:

'If you can take example the results, there were like 20% of interns they can say failed to the OSCEs. So, if we say, those are 20% of people stays in district and prolong their internship, then maybe be 80% who goes to practice. I think this will decrease the number of doctors who are ready to go in practice. But yet, the benefits of that is outweighing the risk of that. So, you can take an example, when I was in internship, it was like a senior clerkship. So we used to get many, many women with infected post caesarean section, but most of them were like poor uterine closure, such kind of things. So, the burden of that is putting pressure to many doctors and occupying them

because maybe one non-ready doctor has gone to district hospital. So, if you can stop that, maybe those complication and their management and their course will be cut, will be cut off. So, I think in many case, using a good way of assessing interns and making sure they..ah, making sure you have ready interns to go there and make sure they are ready to do their job. I think it is still beneficial too rather than letting a good number of them going, but now maybe you are leaving like half of them are ready to do their job and half of them will be causing harm, and they will put the burden to the well-trained doctors to correct their mistakes or medicolegal, such kind of things. I think it is still reasonable.'

e. Assessment can and should be used to influence the training experience

Given the resource-intensive nature of the OSCE and that it seems many gaps in training may need to be addressed before people may feel comfortable introducing a summative assessment for certification purposes, the questions may arise: what is the relative value of continuing to run an OSCE – or otherwise resource-demanding standardized program assessment for that matter - relative to the benefits that it offers?

The comments from examiners suggested that even if not used for certification, the greatest value of the OSCE is the impact it can have on training, both at the individual and programmatic level. For this reason, it is a valuable activity that should continue. Some comments talk about effecting change at the programmatic level:

'…I found that [the OSCE] was a good step to know what we are producing, and…. it can be also be a starting point to look how to ameliorate.'

'Good to know what is going on, where interns doctors are working this will help for further training and for the others who are about to graduate you can plan accordingly…'

'Feedback from this OSCE can help to university and stakeholders improve or upgrade their curriculum.'

'RMDC may use as yardstick for performance, may change policy in training'

'I have no doubt about the results but maybe the question will be, after these results, are some organization working on it so that the internship program can be like reformulated and renewed so that in the local trainers, trainees and how the program can be sufficiently evaluated.'

'I think we…that's why did such assessment, serves as a baseline and we can not use it to pass or to qualify an intern. So, I think that is where it's a…it's a halfway done business that we need to look at in totality.'

Other comments talk about how an exercise like the OSCE can be used to improve the training and learning outcome for the individual trainee:

'…And even if this exam can take another strength….it will increase the way the interns will perform during the internship. Which will increase the responsibility in the internship and then they will lean more than [they are] there. Because they may be doing it as a passive process to go through, but if there's any like additive (inaudible) requirement that they will be asked to respond, they will learn more.'

' …it will be better if this is done not only once a year but at least to be done in the middle and then at the end. This will create kind of stimulation to the intern for more training and more practicing. Thank you.'

'Would be helpful for all interns to do OSCE, clarifies learning in their mind. Would be better if done more than once a year, because they get feedback and learn'

Finally, several examiners comment on their perception that the OSCE exercise would contribute to better patient care and patient safety:

'OSCE for me is a great way of assessing and evaluating medical knowledge of this future general practitioner, thus a good tool of improving quality of management of patients'

'I fully support the OSCE approach as a way to improve knowledge and practical skills of intern doctors'

'The exercise was interesting and fruitful if continued it can be a helpful tool to prepare our interns for better practice in District hospitals'

Discussion

The original intent of this qualitative study was to explore the perspectives and experience of physician examiners on the interns' OSCE in particular and on intern physician assessment more generally. In a commentary on cross-cultural comparisons of assessment of clinical performance, Patel et al. notes, "There is a need to avoid applying instruments of assessment to cultural groups in which proper normative or psychometric research has not been conducted. If a

form of assessment developed in one culture is subsequently applied in another cultural setting, it is imperative to ensure that each application of the test measures the same constructs." (Patel & Agius, 2017)

The previous chapter begins this 'normative or psychometric' research focused on descriptive and psychometric statistics on the Rwandans Interns' OSCE related data. This chapter continues it in a more in-depth, qualitative, exploratory manner from a constructivist paradigm. The findings of both suggest that viewed simply as an assessment instrument in this context, the OSCE and the underlying entrustability-based scoring system used for it was generally viewed as a valid, acceptable and meaningful assessment measure. The present qualitative study however reveals some potential biases in its transfer to the new context and in doing so, identifies some important qualifications or potential threats to validity and acceptability that were not previously recognized. It also further elucidates on cost/feasibility and educational impact.

Assessment Bias and Potential Threats to Validity

Bias and equivalence are two pivotal concepts in the assessment of performance. Bias is said to occur if score differences on the indicators of a particular construct do not correspond to differences in the underlying trait or ability, but are rather attributable to incompatibilities of the underlying constructs, method or items of the assessment with respect to the sample or population being tested. Equivalence is usually accepted to mean the absence of bias. (van de Vijver & Tanzer, 2004) The use and adaptation of assessment instruments in light of this must be considered with respect to cultural validity and specificity in order to optimize the utility of such instruments.(Patel & Agius, 2017) van de Vijver et al. classifies and describes three types of bias (construct bias, method bias and item bias) as typical sources of bias in cross-cultural assessment. The first two come up as relevant in the present study and will be detailed next.

Construct bias can occur if there is only partial overlap in the definitions of the construct across cultures. (van de Vijver & Tanzer, 2004) Generally speaking, careful attention was given to content and construct validity in the process of co-developing the OSCE exam with Rwandan input, as outlined in Chapter 2. Likely as a result, construct bias did not come up as a major issue overall: no examiners expressed that the content did not reflect what would be expected of a competent general medical officer (the underlying construct for the exam), nor did they express that the OSCE failed to capture central elements of this (although it was clear that the OSCE alone should not be used as the only assessment data upon which to base a summative decision). One construct bias that was noted by several examiners was around EPA 13 (demonstrate awareness of one's limitations) and EPA 6 (Provide and receive handover in transitions of care). Although these domains were supported as relevant for assessment when designing the OSCE (see chapter 2), after exam administration it was noted that due to cultural practice norms as well as a lack of easily accessible specialists, in practice this is not yet commonly done. Hence two of the items in Station 6, which required calling a specialist for advice on a complicated patient, may have had poor scores more likely because of this construct bias rather than because of physician lack of ability. In the description of Van de Vijver, this would be an example of differential appropriates of the behaviours associated with the construct (i.e., skills do not belong to the repertoire of the sampled cultural group). (van de Vijver & Tanzer, 2004) It is interesting to note, however, than instead of eliminating the items because of this, the feeling was that a goal of optimal practice would be a cultural practice change towards recognizing one's limitations and calling for help, so it should be maintained for learning purposes even at the risk introducing bias to the assessment scores.

Method bias is further distinguished to include three subtypes: sample bias (occurs when samples used differ in a variety of relevant characteristics other than the target construct),

administration bias (all sources of bias caused by the particular form of administration) and instrument bias (all sources of bias that are associated with the particular assessment instrument). (van de Vijver & Tanzer, 2004) For the OSCE, examiners noted potential sources of bias in all these categories, although the relative impact of these biases varied in their perception. Potential instrument bias was noted as arising from the use of SPs and in particular mannequins, which interns may not have been accustomed to interacting with and/or treating with realism. Potential sources of administration bias were noted, including: language of the exam (suggesting it should be available in French as well as English), SPs not always being consistent in information given, interns not being told that they would have such an exam ahead of time (and therefore being underprepared), anxiety provoked by a timed exam format (which may not reflect day-to-day practice) and the non-summative, low-stakes nature of the exercise (and therefore candidates not taking it seriously/performing to their ability). Some of these potential sources of bias could be more readily addressed than others e.g. translation into French, giving advance warning of exam. In general, these potential sources for bias were seen as minor and there was controversy over how much these biases actually influenced results.

The most pronounced method bias that generated the greatest amount of both consensus and discussion seemed to be around what van de Vijver et al. would deem sampling bias, in the form of incompatibility of the sample caused by differences in education. (van de Vijver & Tanzer, 2004) There was almost unanimous opinion from the examiners, with supportive data from former interns, that deficits in training and supervision of the interns was likely a large underlying factor biasing the results of the assessment, which had demonstrated 'below entrustable' average scores on every station. Whereas the OSCE exam was developed and implemented first in Western regions of the world where the quality of medical training and supervision is often rigorously

controlled by accreditation and evaluation processes, that was not felt to be the case in Rwanda, especially for internship training. Therefore, the sample of candidates (i.e., medical trainees) between the two regions could be considered fundamentally incompatible/incomparable in this regard. This 'sampling bias' results in a threat to the validity of the assessment results. Although the OSCE was measuring the proposed construct of physician competence, the underlying cause explaining the results is an inadequacy of intern physician training in the sample. The presumed lack of intern physician training was viewed as a likely confounder in this context.

A Qualified View of Assessment Acceptability

Above we outlined a potential threat to validity – that is, that results from this assessment exercise can not be completely trusted to be a true reflection of physician competence because of significant deficits in the underlying training of the intern physicians. This in turn seemed to impose an important qualification on its acceptability. Although overall the OSCE appeared to be received as an acceptable and valuable exercise, at the individual trainee level it was seen to be ***most*** acceptable as a formative/learning tool rather than summative/high stakes assessment, until and unless training deficits were addressed. This brings us back to three central themes from the results section:

a. Training, assessment and certification are integrally linked;

b. Training and assessment is a process over time; and

c. Quality assurance for the training and assessment process, and ultimately for certification, depends on program ownership and responsibility.

All of these themes couple assessment together with training, and both are seen as a process occurring over time that must occur with consistent quality and oversight. Assessment will only

be felt to be completely acceptable – particularly for high-stakes decision making – once training is also felt to meet acceptable standards.

This suggests that one cannot entirely trust or accept the results of a clinical performance assessment unless one has established trust in the clinical training process itself. The context of the assessment, which includes the training process, is as important as the assessment tool itself in this respect. This is perhaps the same reason why in some countries, foreign physicians can not simply qualify to become a practicing physician in a new country by challenging and passing an exam. While they may be capable of passing one or even several exams in a new country, the context of their training remains uncertain, likely introducing a fundamental issue of trust in or acceptability of the assessment results until and unless it is verified by equivalency in their training and/or some period of observation in a supervised practice setting. Most organizations in Canada will not allow you to sit for a qualifying exam until some process of training and/or observed practice verification is undertaken. (Medical Council of Canada, 2020; Royal College of Physicians and Surgeons of Canada, 2020a; The College of Family Physicians of Canada, 2020)

Notions of Cross-cultural Acceptability

It is interesting to note that many of the features of trustworthy training and assessment discussed by the focus group parallel concepts espoused by the CBME movement, despite the participants not being explicitly trained in nor particularly aware of CBME principles. The graduated processes of theory to practice with an emphasis on knowledge application, trainee to worker, close supervision to independence, formative to summative assessment with multiple measures and multiple assessors to limit bias and acknowledgement of the potential need for variable time-in-training are all in keeping with CBME. (Caraccio et al., 2002; Carraccio et al., 2016; Kinnear, Warm, & Hauer, 2018) Although the published history and progression of the

CBME movement is largely attributed to and grounded in North America, Western Europe and Australia (Englander et al., 2017), it would seem that quite in contrast to a colonial-like imposition of presumed good ideas, the principles of CBME may find organic resonance in a more global range of cultural contexts. While the transplantation of a particular assessment tool or assessment system from one culture to another may not demonstrate equivalence, it is possible that principles of training *and* assessment, taken and integrated together with appropriate care and resources, may still have cross-cultural applicability.

In addition to CBME principles noted above, another example of such cross-cultural acceptability in the present context was the concept of entrustability and the related entrustability-based scoring scale that was used for as the basis for assessment in the interns' OSCE. This particular concept was defined for, presented to, and applied by examiners during their examiner training based almost on entirely on Western/European ideology and experience (Englander & Carraccio, 2014; Ten Cate et al., 2016) Although checklists were provided as guides to observation, it was explicitly emphasized that ultimately it was examiner *judgment* – not a strictly applied checklist and/or tally-count – that would determine the score for any given item. In this way, it perhaps struck a good balance between the practice-based value judgements that served as the basis for assessment as outlined in a cross-cultural assessment study in the Middle East by Wilbur et al. (Wilbur, Hassaballa, Mahmood, & Black, 2017) versus "…objective observations against clear descriptors or standards rather than value judgements…" that Patel et al. suggest should be the goal for assessment in any cultural setting. (Patel & Agius, 2017) Regardless of the original intention of this hybrid approach (i.e. criterion-guided, judgement based, entrustability-anchored scoring), it was of interest to explore in the focus group how it actually played out in practice.

Focus group comments suggested relative ease and liking for the use of this entrustability-based scoring approach and no one indicated confusion or frustration with its application. Some individuals spoke of it more intuitively as a general feeling about an observed performance, in contrast to others who felt that the defined criteria allowed for an objective breakdown of the score that could then be translated into a judgment. Regardless of the perspective, the process of coming to an end-decision seemed equivalently comfortable and meaningful for both. It was noted that a pre-discussion among experts of what criteria might be most important for entrustable performance for a given station or EPA was felt to be helpful in creating a common understanding upon which they could base their assessments.

The importance of program ownership and responsibility as well as ongoing quality assurance for the physician training and assessment process is of fundamental importance, as evidenced both by focus group themes in the Rwandan setting as well as in Western medical education literature and practice. The CBME movement in the West recently outlined this at the local, institutional and national levels in its 'charter for clinician-educators' asking for commitment: to supervision that balances patient safety with professional development of learners, to the effectiveness and efficiency of assessment strategies, to workplace assessment and programs of evaluation, to faculty development and collaboration of all stakeholders to achieve vertical and horizontal integration. (Carraccio et al., 2016) That this will require significant investment and work is clear from their language, even in resource-rich settings with a history of multiple funded independent education and regulatory bodies to promote, facilitate and monitor implementation i.e. American Association of Medical Colleges (AAMC), Accreditation Council for Graduate Medical Education (ACGME), Association of Faculties of Medicine of Canada (AFMC), Royal

College of Physicians and Surgeons of Canada (RCPSC), College of Family Physicians of Canada (CFPC) .

The recommendations for the Rwandan context, based on the findings of our study, are not necessarily different: program ownership and responsibility as well as ongoing quality assurance for both the training and assessment process at both the national and local levels. However, the starting point is different and therefore, the focus and scope of activities may also be different. Clarity and definition of the role of the intern physician in the medical system, of training and supervision expectations and of what constitutes adequate clinical exposure should come first. A program of meaningful formative and summative assessment, using multiple methods and multiple assessors, as well as well-defined certification standards (actual or potential) should likely come next. Going through these steps in a collaborative manner with the various stakeholders involved (Ministry of Health, RMDC, District hospitals, intern physicians and potentially new partners like the University of Rwanda medical school, Ministry of Education) to establish common goals and understanding will likely be helpful in facilitating success in implementation. Clear assignment of responsibilities – which institution or organization is responsible for providing what component(s) of the defined training and assessment process – needs to be delineated. Investment in faculty development and leadership at the local level will also be a critical step. Finally, regular periodic monitoring and accreditation procedures lead at the national level while closely involving the local level will be important for ensuring quality and accountability.

Managing Cost and Feasibility

A multitude of resource requirements for effective training and assessment were discussed in the focus group and are outlined in Table 14 under the general categories of time, people, purpose, expertise and training, and money. There were comments suggesting that a lack of money

should neither the biggest nor the most prohibitive barrier to reaching the end goal. There may be strategic, resourceful ways of harnessing people, time and common purpose to achieve a vision for better quality training and assessment, despite a low-resource setting. There was a generally positive outlook around how the requirements could be managed and the feeling was that the benefits offered outweighed the costs. Still, time and people are finite, at least at a particular moment in time, and this presented some tough realities. There are no extra physicians that can be easily introduced to help balance the patient care – supervision demand, and for physicians that are already overwhelmed with clinical demand, asking them to take on additional supervisory responsibilities may simply not be realistic. Although specialists are now being 'posted' by the Ministry of Health to specific district hospital sites for a set term upon graduation, there is high turnover of specialists (and sometimes of general medical offers) at these peripheral centres that poses a continuous challenge to the sustainability and success of physician training programs.

Harnessing Educational Impact

A final major theme identified was that assessment can and should be used to influence the training experience. Although the OSCE may not necessarily be universally accepted for the purpose of a high-stakes summative or certification assessment in the Rwandan context at present, several comments suggest that its greatest value is the impact it can have on influencing the training process and ultimately improving training outcomes, at the individual, programmatic and patient care/patient safety levels. In this way, the focus of the assessment moves away from assessment *of* learning of the individual learner, to assessment *for* learning for both the individual learner as well for the training program. (Martinez & Lipson, 1989; Schuwirth & Van der Vleuten, 2011b) This shift of focus resonates with what Mumtaz et al. previously suggested in the context of cross-cultural comparisons of assessment, "…that the importance of holistic educational

supervision, which would be applicable across all cultural groups, ought to be given a higher profile. This has broader implications than assessment systems transplanted from one culture to another". (Patel & Agius, 2017)

Conclusion

In summary, a qualitative exploration of the perspectives and experience of examiners on postgraduate physician assessment in a culturally different, limited-resource setting revealed a number of themes as summarized below and graphically shown in Figure 4:

a. Training, assessment and certification are integrally linked;

b. Training and assessment is a process over time;

c. Quality assurance for the training and assessment process, and ultimately for certification, depends on program ownership and responsibility;

d. Quality has requirements and costs; and

e. Assessment can and should be used to influence the training experience.

As trust in the assessment is predicated on trust in the training process, in this new setting where trust in the training process is questionable, the validity of the assessment is threatened. It is this context-specific threat to validity, rather than an inherent cultural values based rejection of Western training or assessment principles per se, that influences and mitigates the acceptability of the assessment. With validity and acceptability thus affected, the utility of a 'transplanted' OSCE is different than what it may be in its native/originator context. In particular, such an assessment takes on a more formative/learning rather than summative or certification role at the individual level and importantly, it should act as both a driver and a marker of change at the programmatic level.

Chapter 4: Conclusion

Revisiting the Cross-cultural Utility of an Assessment

Recall that we had proposed to apply the following conceptual equation to determine the utility of an assessment method for professional competence (C.P. Van Der Vleuten, 1996):

$$\text{Assessment Utility} = V * R * E * C * A$$

(V=validity, R= reliability, E =Educational impact, C=Cost and A=Acceptability)

The suggestion is that the utility of an assessment is the multiplicative function of these variables with different weights (w) associated with each of them. As there is no one perfect assessment, perfect utility is unattainable. In practice, we will always need to compromise and assign different weights in different individual situations. What is important to note is that if any of the elements is zero, then the utility will be zero.

In the case of introducing the Western-derived OSCE assessment to Rwanda, we applied a mixed methods approach, using both quantitative/numeric data (reported mainly in Chapter 2) as well as qualitative/narrative data (reported mainly in Chapter 3) to explore each of the above utility variables. Our purpose in doing so was to conduct a comprehensive and meaningful evaluation of the 'transplant' of a Western-derived evaluation tool into the culturally and economically different setting of Rwanda. Our findings will be reviewed below.

Divergence versus Convergence of Findings

Part of the rationale in including both quantitative and qualitative approaches was to see if these two methods would result in convergent or divergent findings, as well as to perhaps provide better insight as to why the case was so. As will be detailed below, it would seem that while our findings from quantitative and qualitative components share some convergence when characterizing VRECA variables, there is important divergence in findings particularly on the

variables of validity and acceptability of the OSCE in this new setting. The sections that follow below further describe these trends and elaborate on the reasons why.

Validity

Our quantitative analysis of data collected both before and after the OSCE (presented mainly in Chapter 2) demonstrated compelling evidence of content and face validity. Care taken during the creation of the OSCE ensured >75% of experts felt content was of high relevance and priority and that it would be tested in a realistic manner. After the OSCE, 100% of examiners agreed that the OSCE content was relevant to the future clinical practice of intern doctors and that the content of their specific station as relevant and realistic to clinical practice.

The qualitative analysis however reviewed some important nuances and potential threats to validity. Looking at potential for bias in the 'export' of this tool, based largely on focus group comments, there was some suggestion of *construct bias* (where there is only partial overlap in the definitions of the construct across cultures) and *method bias* (further distinguished into *sample bias*, *administration bias* and *instrument bias*). (van de Vijver & Tanzer, 2004) The most pronounced of these, and the one that generated the greatest discussion and agreement from examiners, was sampling bias in the form of incompatibility of the sample caused by differences in education between Western vs Rwandan populations. (van de Vijver & Tanzer, 2004) There was almost unanimous opinion from the examiners, with supportive data from former interns, that deficits in training and supervision of the interns was likely a large underlying factor biasing the results of the assessment, which had demonstrated 'below entrustable' average scores on every station. Whereas the OSCE exam was developed and implemented first in Western regions of the world where the quality of medical training and supervision is often rigorously controlled by accreditation and evaluation processes, that was not felt to be the case in Rwanda, especially for

internship training. This 'sampling bias' results in a threat to the validity of the assessment results. Although the OSCE was measuring the proposed construct of physician competence, the underlying cause explaining the results is an inadequacy of intern physician training in the sample. The presumed lack of intern physician training was viewed as a likely confounder in this context.

Reliability

Reliability typically is defined by the following general equation (Streiner et al., 2015):

$$Reliability = \frac{Subject\ variability}{Subject\ variability + Measurement\ error}$$

There are several ways to measure reliability, all based on numeric or quantitative data. In the case of the OSCE, the calculation of Cronbach's α, best described as a measure of internal consistency (i.e., do all the items in the test or assessment measure the same construct), was applied. Based on using all obtained station scores (51 items), the overall reliability the OSCE was calculated at Cronbach's α of 0.90. It has been suggested that for high-stakes examinations, a reliability index such as a Cronbach's α or g-coefficient of greater than 0.7 or 0.8 is necessary (Harden et al., 2016). Thus, the reliability as measured in this case would suggest that this OSCE could be suitable for use for a high-stakes summative or licensing examination. However, threats to validity of the examination influence the acceptability of the assessment for such a purpose, as we shall discuss next.

Acceptability

In terms of acceptability, data from Chapter 2 suggests that the OSCE was perceived as a useful, relevant and realistic learning/assessment tool by the significant majority of examiners. A clear majority felt it should be run again in future years and would be willing to be an examiner again. Examiners generally felt their station content was of appropriate difficulty level, time allotment and relevance. This is important because besides the validity, reliability, cost and

feasibility of an assessment tool, the acceptability likely strongly influences is uptake and use going forward (C.P. Van Der Vleuten, 1996). Data similarly supports the acceptability of the OSCE from the examinees' perspective, with greater than 90% of intern physicians indicating it was a useful learning experience that should be run for future interns, and nearly 80% endorsing it as a method of assessment. It is interesting to note that the question with the greatest distribution of opinion was around whether the OSCE could be used as part of a licensing process or standard in the future.

In the qualitative analysis of Chapter 3, the reasons behind why acceptability of the OSCE as a high-stakes summative standard-setting or licensing exam was questionable, despite evidence of high content/face validity, reliability and acceptability from Chapter 2, became more clear. Above we outlined a potential threat to validity – that is, that results from this assessment exercise can not be completely trusted to be a true reflection of physician competence because of significant deficits in the underlying training of the intern physicians. This in turn seemed to impose an important qualification on its acceptability. Although overall the OSCE appeared to be received as an acceptable and valuable exercise, at the individual trainee level it was seen to be ***most*** acceptable as a formative/learning tool rather than summative/high stakes assessment, until and unless training deficits were addressed.

Educational Impact

The potential for educational impact, defined as the ability of an assessment to influence the learning of the individual, or the curricular design of the learner program for the institution (C.P. Van Der Vleuten, 1996), was best estimated by the qualitative analysis from Chapter 3. Although the OSCE may not necessarily be universally accepted for the purpose of a high-stakes summative or certification assessment in the Rwandan context at present, several comments

suggest that its greatest value is the impact it can have on influencing the training process and ultimately improving training outcomes, at the individual, programmatic and patient care/patient safety levels. In this way, the focus of the assessment moves away from assessment *of* learning of the individual learner, to assessment *for* learning for both the individual learner as well for the training program. (Martinez & Lipson, 1989; Schuwirth & Van der Vleuten, 2011b) It is important to note that we say that we *estimated the potential* for educational impact rather than *measured the actual* educational impact. This is because as noted previously, educational impact is likely best measured and tracked over time, either at an individual or programmatic level. The limited time-horizon/follow-up period to the study, study scope/resource limitations and as well as the 'exit-timed' nature of the OSCE which occurred at the end of internship for the individual learner, meant that it was not favourably designed to accurately or precisely capture long-term actual educational impact.

Cost/Feasibility

The approximate cost of the OSCE as reported in Table 4 was about 11 million RWF ($12,000 USD) per year, about $250 USD per examinee. This was inclusive of all examiner and facilitator/SP training sessions. This appears to be on the lower end of the wide cost-range ($11 to $1200 USD per candidate) for OSCEs as reported in the literature (Harden et al., 2016), but still not insignificant in the context of the Rwandan limited-resource setting.

This $250 USD per examinee figure is best seen as a rough estimate of cost. There is a wide variation in the way that costs are calculated and reported in the literature, often resulting in a 'high-end cost' and 'low-end cost' (Reznick et al., 1993). Our study was no different with respect to this variation. It is important to mention some elements that in this case may contribute to either a 'high-end' or 'low-end' reporting. Most notably, the reported cost did not include any

cost/payment for three key OSCE leadership personnel: OSCE lead, Simulation specialist, Administrator/Data manager. These costs are difficult to estimate for a number of reasons: their specific hours were not systematically tracked, their hypothetical local hourly rate or payscales are not defined, the demands of their roles were variable year-to-year (often more involved in first-iteration) and their actual cost would probably change based on whether or not these were stand-alone jobs (unlikely) versus responsibilities integrated into a larger job description (more likely). It also did not include costs of the actual vehicle that helped with transport of materials between sites (as this was a TSAM project owned vehicle). It also did not include the costs associated with development and piloting of the OSCE stations. Conversely, the reported cost did include all costs related to the many training sessions for the OSCE. If this was to become a routine year-upon-year practice, then with sufficient years of institutional memory, such extensive training may no longer be required. Table 4 also detailed the human resources, physical space and equipment requirements required and included as part of the monetary cost. Successful execution of the OSCE for two iterations using mostly locally-sourced resources offers some persuasive evidence of feasibility.

Qualitative exploration of cost and feasibility issues highlighted the time, people, expertise and training and money that may be required for effective training *and* assessment, which was felt to go hand-in-hand. A key message seemed to suggest that neither money nor a limited resource environment should be the biggest nor the most prohibitive barrier to reaching what they felt was a worthwhile end goal of improved training and assessment. There may be strategic, resourceful ways of harnessing people, time and common purpose to achieve a vision for better quality training and assessment, despite a low-resource setting. There was a generally positive outlook around

how the requirements could be managed and the feeling was that the benefits offered outweighed the costs.

Insights on Evaluating the Utility of Assessment

In summary from the above, it would appear that there is convincing evidence of validity, reliability, acceptability and cost/feasibility, particularly based on objective/numerical quantitative and psychometric measures presented in Chapter 2. Subjective and narrative data analysis from a constructivist perspective revealed important context-specific limitations particularly on validity, which then appeared to influence acceptability. Analysis of this narrative data was also the best way to gauge educational impact, which in the present context seemed to emphasize assessment *for* learning for both the individual learner as well for the training program over assessment *of* learning of the individual learner. (Martinez & Lipson, 1989; Schuwirth & Van der Vleuten, 2011b)

It is worth noting that it seemed to be threats to validity of the assessment (identified as inadequate training of interns) that limited its acceptability (i.e. hesitance in applying it as a high-stakes or summative assessment in Rwanda, despite demonstrating high quantitative measures of content/face validity and reliability). It was not an inherent cultural rejection. The OSCE was not dismissed by the belief system of the Rwandan medical professionals as an assessment tool that lacked utility in their setting because of a fundamental difference in cultures, but rather it was the threat to validity that influenced the acceptability. In this way, we can see how perceived validity can place an upper limit on the acceptability of an assessment.

It is also important to note that had we relied exclusively on traditional quantitative and psychometric approaches to assessing VRECA, we would have missed some important information that is directly relevant to VRECA and immediately influenced the assessment utility.

This highlights the importance of using mixed-methods approaches (i.e., both quantitative and qualitative) in evaluating the components of utility and revisiting components of VRECA both before and after implementation. Some of its components (i.e., acceptability, educational impact) may be better suited to characterization or 'measurement' with qualitative inquiry over quantitative measures, and other components (i.e. validity, acceptability) may change from what was established (often theoretically) before implementation to what exists after implementation. The same likely applies to implementing new assessments or assessment approaches even within the same culture. In Canada for example, sweeping changes have recently been (and continue to be) implemented in the postgraduate education field in accordance with principles of CBME, which is has been branded as 'Competency by Design' (CBD) by the RCPSC. The specialty of anesthesia initially formulated its first set of national EPAs, which would serve as their units/items of assessment, in 2015-16 with a first implementation in 2017. They have since revisited their EPAs after implementation and found generally that content validity was mostly re-affirmed, but acceptability and feasibility were both challenges. There were too many EPAs and too many required assessments to be practical or acceptable to trainees and faculty. As a result, several revisions were made towards the goal of reducing the total number of EPAs and/or observed assessments required, as well as improving their wording to make them more accessible. (Vergel de Dios, 2020)

The above phenomenon, underscoring the importance of both quantitative and qualitative characterization of validity in particular, highlights the need to appropriately conceptualize validity. The call for this has been made elsewhere in old literature pre-dating CBME, as well as recent literature regarding assessment in the era of CBME. The validity of an assessment, particularly a summative assessment, is of paramount concern. The tradition in the 20th century

was the adoption of a systematic approach to the rigorous interrogation of assessment data in order to determine the accuracy of a judgment. However, it is now understood that traditional representations of validity (e.g., numeric measurements of content, criterion and construct validity) can result in a limited and superficial understanding of the accuracy of a judgment. More than 25 years ago, Messick proposed a definition of validity that moved past the statistical accuracy of quantitative scores: "Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." By this conception, validity is not a "number" but, rather, an argument that supports the final judgment about "true" physician competence. (Harris et al., 2017) In the case of Rwanda, the numbers alone suggested that the validity of the test and the results were assured. However, the information derived from the focus group held after implementation highlighted important threats to the argument for validity due to specific contextual factors, in particular, that of significant gaps in training of the intern physicians.

Just as traditional concepts of validity have evolved, so too has reliability in the CBME era. Hodges pointed out that the notion of *subjectivity* had taken on the connotation of *bias*, and that standardization was touted as the ticket to reliability, even though adequate sampling mitigates bias and is the main determinant of reliability. One can have objective measures (such as standardized checklists) that yield unreliable scores, and subjective measures (such as expert judgments using global rating scales) that provide reliable scores. (Lockyer et al., 2017) In our case study, this is perhaps evidenced by the fact that all our scores were entrustability-based global rating scales. Although checklists were provided for guidance, they were not strictly applied in any sort of tally or prescriptive fashion when it came to the examiner giving a score – it was

ultimately based on their judgment. Even with this 'subjective' approach, our reliability index (Cronbach's α) was a respectable 0.90. There could be several reasons for this, however the adequate sampling (multiple stations, multiple assessors) as well as examiner training was likely contributory.

Even if the reliability calculation had not measured high, in recent years the importance of reliability has been somewhat de-emphasized in recognition of the limitations of reliability metrics and the relative paramount importance of validity. Central to this is that reliability indices such as the Cronbach's alpha (based on classical test theory) or g-coefficient (based on generalizability theory) aim to describe how much of the observed variance in a given sample comes from true between-subject differences vs random-error (e.g., raters, items, sites, test execution, etc.), as suggested by the below general form of the reliability equation (Streiner et al., 2015):

$$Reliability = \frac{Subject\ variability}{Subject\ variability + Measurement\ error}$$

It is important to note then that if the true baseline between-subject variability is low, as may be the case for highly trained physicians, the reliability index will remain low regardless of accurate measurement. Equally important to consider is the relevance of reliability to the assessment goal. If the goal of the assessment is to determine if candidate A is better or different then candidate B, then an instrument that has a high ability to discriminate between subjects becomes important. If however the main goal is to determine whether or not candidate A and candidate B meet a certain standard, as is increasingly the case in CBME, then the ability to discriminate between candidates – which is the reliability metric in the sense of being able to isolate and measure between-subject differences – becomes less relevant. What becomes more relevant is convincing evidence that your assessment measures what it intends to measure (i.e. validity) and that appropriate standard-setting is undertaken. For these reasons, it has been acknowledged that validity should not be

sacrificed for reliability and feasibility (Harden et al., 2016) and that a test with high reliability is not always better than a test with a lower reliability (Schuwirth & van der Vleuten, 2011a).

Cultural Common Ground in Principles of CBME and Entrustability

As noted in Chapter 3, it was remarkable that in the course of the examiners' focus group discussion that started as an exploration on assessment, the discussion quickly and repeatedly turned to trustworthy training *and* assessment. In doing so, many of the features of trustworthy training and assessment discussed by the focus group paralleled concepts espoused by the CBME movement, despite the participants not being explicitly trained in nor particularly aware of CBME principles. The graduated processes of theory to practice with an emphasis on knowledge application, trainee to worker, close supervision to independence, formative 'continuous evaluation' to summative assessment with multiple measures and multiple assessors who have received some training and acknowledgement of the potential need for variable time-in-training, all of which are themes that came out of the focus group, are all in keeping with CBME. (Caraccio et al., 2002; Carraccio et al., 2016; Kinnear et al., 2018; Lockyer et al., 2017) Although the published history and progression of the CBME movement is largely attributed to and grounded in North America, Western Europe and Australia (Englander et al., 2017), it would seem that quite in contrast to a colonial-like imposition of presumed good ideas, the principles of CBME may find organic resonance in a more global range of cultural contexts. While the transplantation of a particular assessment tool or assessment system from one culture to another may not demonstrate equivalence, it is possible that principles of training *and* assessment, taken and integrated together with appropriate care and resources, may still have cross-cultural applicability.

The close coupling of assessment and training has received greater attention in CBME. Two fundamental and yet essentially different rationales are *assessment of learning* and *assessment for learning*. Before the introduction of CBME, the former was emphasized; however as CBME becomes established, the focus is shifting *to assessment for learning*. Van der Vleuten et al. suggest that "whenever assessment becomes a goal in itself, it is trivialized and will ultimately be abandoned. Assessment has utility insofar as it succeeds in driving learning, is integrated in a routine and ultimately comes to be regarded as indispensable to the learning practice." Thus, if the primary purpose in assessment in CBME is to drive learning, and our secondary purpose is to make judgments about readiness to progress, assessment programs need to be designed accordingly. (Lockyer et al., 2017) Stated another way, the increasing importance of *assessment for learning* can be seen as a relatively heavy weighting of educational impact in the utility equation. In the Rwandan context, the primary value of the OSCE assessment was seen in its educational impact or *assessment for learning* both at the individual and programmatic level, although it was acknowledged that at some point, after training deficits have been addressed, then *assessment of learning* also becomes a legitimate secondary goal. *Assessment of learning* aligns with the continuing need to gauge progress against targeted outcomes and criterion-referenced standards in CBME. (Lockyer et al., 2017)

In addition to CBME principles discussed above, another example of cross-cultural acceptability in the present context was the concept of entrustability and the related entrustability-based scoring scale that was used for as the basis for assessment in the interns' OSCE. This particular concept was defined for, presented to, and applied by examiners during their examiner training based almost on entirely on Western/European ideology and experience (Englander & Carraccio, 2014; Ten Cate et al., 2016) Although checklists were provided as guides to

observation, it was explicitly emphasized that ultimately it was examiner *judgment* – not a strictly applied checklist and/or tally-count – that would determine the score for any given item. Focus group comments suggested relative ease and liking for the use of this entrustability-based scoring approach and no one indicated confusion or frustration with its application. Some individuals spoke of it more intuitively as a general feeling about an observed performance, in contrast to others who felt that the defined criteria allowed for an objective breakdown of the score that could then be translated into a judgment. Regardless of the perspective, the process of coming to an end-decision seemed equivalently comfortable and meaningful for both. It was noted that a pre-discussion among experts of what criteria might be most important for entrustable performance for a given station or EPA was felt to be helpful in creating a common understanding upon which they could base their assessments.

A Global Perspective in the Challenges of Implementing CBME

There appears to be interest in the uptake of CBME principles in many different global settings. In the setting of Rwanda, based on a sampling of physician educators and trainees who were involved in the OSCE, there seems to be organic support for many of the CBME principles both in training and assessment. The Postgraduate Medical Education WFME Global Standards 2015 similarly reflect a general shift from a time-and-process based training to competency-based training, and several of their 'basic' (i.e., minimum) as well as quality development (i.e., optimal) standards are in keeping with principles of CBME. (World Federation for Medical Education, 2017) Many of the WFME standards are also in keeping with other recommendations about governance and leadership, programme evaluation and selection/development of trainers that were identified as lacking in the Rwandan internship program. The results of our findings from our

participants in Rwanda, as well as the standards published by the WFME, suggest the imperative for changes in training and assessment in tandem, and not just one or the other in isolation.

Many perceived gaps in the training and assessment process of the Rwandan internship were identified in the course of our research. These include deficiencies in: practical application of knowledge, appropriate supervision, understanding of expectations and limitations of intern as a trainee versus worker, meaningful formative and summative assessment leading to certification standards, adequate clinical exposure and flexible duration of training. There was recognition from the examiners that the ownership and responsibility of quality assurance for training and assessment is shared at both the local level, in the form of hospitals that are hosting internship programs as well as the national level, in the form of the Rwanda Medical and Dental Council (RMDC) that is administering the internship program. There are several other stakeholders, including potentially the Ministry of Health, the Ministry of Education and the University of Rwanda College of Medicine and Health Sciences, that would also have a role to play in the continuous quality improvement of the internship program.

Caverzagie et al. recently outlined what they see as overarching challenges to the implementation of CBME. They included the following (Caverzagie et al., 2017) :

1. Aligning regulatory stakeholders to support competency-based education and training.
2. Integrating educational and clinical redesign efforts to align curricular objectives with experiential training.
3. Establishing defined outcomes that reflect the needs of patients and populations in which individuals, programs and institutions can be measured.
4. Ensuring accountability among all stakeholders for the achievement of defined outcomes.

While these challenges were identified more likely based on the Western experience (given the distinctly North American and Western European authorship) and certainly not written with the specific cultural and limited-resource context of Rwanda in mind, based on our research findings they still seem applicable to the Rwandan context. The precise nature and magnitude of these challenges may differ between Western vs other global settings, as well as the resources of people, time and finances that are readily available to address them. However, in the view of Rwandans, this should not discourage or prevent strategies and initiatives that work towards the implementation of competency-based training and assessment for the Rwandan internship program.

List of References

Abdelaziz, A., Hany, M., Atwa, H., Talaat, W., & Hosny, S. (2016). Development, implementation, and evaluation of an integrated multidisciplinary Objective Structured Clinical Examination (OSCE) in primary health care settings within limited resources. *Med Teach, 38*(3), 272-279. doi:10.3109/0142159X.2015.1009018

Al-Chalabi, T., Al-Na'Ama, M., Al-Thamery, D., Alkafajei, A., Mustafa, G., Joseph, G., & Sugathan, T. (1983). Critical performance analysis of rotating resident doctors in Iraq. *Medical Education*(17), 378-384.

Association of American Medical Colleges. (2014). *Core Entrustable Professional Activities for Entering Residency: Faculty and Learners Guide,* . Retrieved from https://members.aamc.org/eweb/upload/core%20EPA%20Curriculum%20Dev%20Guide.pdf

Association of Faculties of Medicine of Canada. (2016). *AFMC Entrustable Professional Activities for the Transition from Medical School to Residency*. Retrieved from Canada: https://afmc.ca/sites/default/files/documents/AFMC_Entrustable_Professional_Activities_EN_0.pdf

Bakir, I., & Abdel-Razig, S. (2019). The Internship Year: A Potential Missed Opportunity to Expand Medical Access in International Settings. *Journal of Graduate Medical Education, 11*(4s), 30-33. doi:10.4300/jgme-d-19-00117

Barman, A. (2005). Critiques on the Objective Structured Clinical Examination. *Annals of the Academy of Medicine, Singapore, 34*(8), 478-482.

Barzansky, B. P. (2010). Abraham Flexner and the Era of Medical Education Reform. *Academic Medicine, 85(9) Supplement, A Snapshot of Medical Student Education in the United States and*(Canada), Reports from 128 Schools:S119-S125.

Bleakley, A., Brice, J., & Bligh, J. (2008). Thinking the post-colonial in medical education. *Med Educ, 42*(3), 266-270. doi:10.1111/j.1365-2923.2007.02991.x

Caraccio, C., Wolfsthal, S., Englander, R., Ferentz, K., & Martin, C. (2002). Shifting Paradigms: from Flexner to Competencies. *Academic Medicine, 77*(5), 361-367.

Carraccio, C., Englander, R., Van Melle, E., Ten Cate, O., Lockyer, J., Chan, M. K., . . . International Competency-Based Medical Education, C. (2016). Advancing Competency-Based Medical Education: A Charter for Clinician-Educators. *Acad Med, 91*(5), 645-649. doi:10.1097/ACM.0000000000001048

Caverzagie, K. J., Nousiainen, M. T., Ferguson, P. C., Ten Cate, O., Ross, S., Harris, K. A., . . . Collaborators, I. (2017). Overarching challenges to the implementation of competency-based medical education. *Med Teach, 39*(6), 588-593. doi:10.1080/0142159X.2017.1315075

David Hodges, B. M. D. P. (2010). A Tea-Steeping or i-Doc Model for Medical Education? [Miscellaneous]. *Academic Medicine, 85(9) Supplement, A Snapshot of Medical Student Education in the United States and*(Canada), Reports from 128 Schools:S134-S144.

Day, S. H., & Nasca, T. J. (2019). ACGME International: The First 10 Years. *Journal of Graduate Medical Education, 11*(4s), 5-9. doi:10.4300/jgme-d-19-00432

De Almeida Troncon, L. E. (2004). Clinical skills assessment: limitations to the introduction of an "OSCE" (Objective Structured Clinical Examination) in a traditional Brazilian medical school. *Sao Paulo Med J, 122*(1), 12-17.

Englander, R., & Carraccio, C. (2014). *Entrustable Professional Activities as an Organizing Framework for Assessment across the Contimuum.* Paper presented at the International Conference on Residency Education (ICRE), Toronto, Canada.

Englander, R., Frank, J. R., Carraccio, C., Sherbino, J., Ross, S., Snell, L., & Collaborators, I. (2017). Toward a shared language for competency-based medical education. *Med Teach, 39*(6), 582-587. doi:10.1080/0142159X.2017.1315066

General Medical Council. (2013). Good Medical Practice. Retrieved from http://www.gmc-uk.org/guidance/

Harden, R., Lilley, P., & Patricio, M. (2016). *The Definitive Guide to the OSCE*. Edinburgh: Elselvier Ltd.

Harris, P., Bhanji, F., Topps, M., Ross, S., Lieberman, S., Frank, J. R., . . . Collaborators, I. (2017). Evolving concepts of assessment in a competency-based world. *Med Teach, 39*(6), 603-608. doi:10.1080/0142159X.2017.1315071

Hays, R. (2014). The potential impact of the revision of the Basic World Federation Medical Education Standards. *Med Teach, 36*(6), 459-462. doi:10.3109/0142159X.2014.907881

Kinnear, B., Warm, E. J., & Hauer, K. E. (2018). Twelve tips to maximize the value of a clinical competency committee in postgraduate medical education. *Med Teach, 40*(11), 1110-1115. doi:10.1080/0142159X.2018.1474191

Lockyer, J., Carraccio, C., Chan, M. K., Hart, D., Smee, S., Touchie, C., . . . Collaborators, I. (2017). Core principles of assessment in competency-based medical education. *Med Teach, 39*(6), 609-616. doi:10.1080/0142159X.2017.1315082

Martinez, M., & Lipson, J. (1989). Assessment for Learning. *Educational Learning, 46*(7), 73-75.

Medical Council of Canada. (2020). How to Become a Practising Physician In Canada. Retrieved from https://physiciansapply.ca/how-to-become-a-practising-physician-in-canada/

Miller, G. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9 Suppl), S63-67.

Moiz, B., Ali, S. K., Rashid, A., Shariq, M., & Karim, F. (2019). Development and Pilot Testing of a Novel Tool for Evaluating Practical Skills in Hematopathology Residents in Pakistan. *Journal of Graduate Medical Education, 11*(4s), 177-180. doi:10.4300/jgme-d-18-00361

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., . . . Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach, 33*(3), 206-214. doi:10.3109/0142159X.2011.551559

Pailthorpe, B. (2017). Emergent Design.

Patel, M., & Agius, S. (2017). Cross-cultural comparisons of assessment of clinical performance. *Med Educ, 51*(4), 348-350. doi:10.1111/medu.13262

Republic of Rwanda MInistry of Health. (2011). *Medical Intern Handbook, Information and Logbook for Medical Interns*. Rwanda: Republic of Rwanda, Ministry of Health.

Republic of Rwanda Ministry of Health. (2020). Key Health Indicators. Retrieved from https://moh.gov.rw/index.php?id=514

Reznick, R. K., Smee, S., Baumber, J. S., Blackmore, D., & Berard, M. (1993). Guidelines for Estimating the Real Cost of an Objective Structured Clinical Examination. *Academic Medicine, 68*(7), 513-517.

Richards, L. (2015). *Handling Qualitative Data, A Practical Guide, Third Edition.* (Third Edition ed.): SAGE Publications.

Royal College of Physicians and Surgeons of Canada. (2020a). International Medical Graduates. Retrieved from http://www.royalcollege.ca/rcsite/credentials-exams/assessment-international-medical-graduates-e

Royal College of Physicians and Surgeons of Canada. (2020b). Royal College International. Retrieved from http://www.royalcollege.ca/rcsite/international-old-e

Sainterant, O., Clisbee, M., & Julceus, E. F. (2019). Introducing the Objective Structured Clinical Examination in Haiti. *Journal of Graduate Medical Education, 11*(4s), 199-200. doi:10.4300/jgme-d-19-00224

Sasaki, H., Archer, J., Yonemoto, N., Mori, R., Nishida, T., Kusuda, S., & Nakayama, T. (2005). Assessing doctors' competencies using multisource feedback: validating a Japanese version of the Sheffield Peer Review Assessment Tool (SPRAT). *BMJ Open, 2015*(5). doi:10.1136/bmjopen-2014-
007135

Scheele, F., & Ten Cate, O. (2007). Viewpoint: Competency-Based Postgraduate Training: Can We Bridge the Gap between Theory and Clinical Practice? *Academic Medicine, 82*(6), 542-547.

Schuwirth, L. W., & van der Vleuten, C. P. (2011a). General overview of the theories used in assessment: AMEE Guide No. 57. *Med Teach, 33*(10), 783-797. doi:10.3109/0142159X.2011.611022

Schuwirth, L. W., & Van der Vleuten, C. P. (2011b). Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach, 33*(6), 478-485. doi:10.3109/0142159X.2011.565828

Schwarz, M. R., Wojtczak, A., & Stern, D. (2007). The outcomes of global minimum essential requirements (GMER) pilot implementation in China. *Med Teach, 29*(7), 699-705. doi:10.1080/01421590701749821

Stern, D. T., Ben-David, M. F., De Champlain, A., Hodges, B., Wojtczak, A., & Schwarz, M. R. (2005). Ensuring global standards for medical graduates: a pilot study of international standard-setting. *Med Teach, 27*(3), 207-213. doi:10.1080/01421590500129571

Stern, D. T., Wojtczak, A., & Schwarz, M. R. (2003). The assessment of global minimum essential requirements in medical education. *Medical Teacher, 25*(6), 589-595. doi:10.1080/0142159032000151295

Stillman, P., Wang, Y., Ouyang, Q., Zhang, S., Yang, Y., & Sawyer, W. (1997). Teaching and assessing clinical skills: a competency-based programme in China. *Medical Education, 31*(1), 33-40.

Streiner, D., Norman, G., & Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to their development and use.* (5th Edition ed.). United Kingdom: Oxford University Press.

Talib, Z., Narayan, L., & Harrod, T. (2019). Postgraduate Medical Education in Sub-Saharan Africa: A Scoping Review Spanning 26 Years and Lessons Learned. *Journal of Graduate Medical Education, 11*(4s), 34-46. doi:10.4300/jgme-d-19-00170

ten Cate, O. (2006). Trust, competence and the supervisor's role in postgraduate training. *BMJ, 333*(7571), 746-748. doi:10.1136/bmj.38961.475718.68

ten Cate, O. (2015, Saturday February 28, 2015). *Entrustable Professional Activities as a Framework for the Assessment of Residents.* Paper presented at the The 2015 ACGME Annual Educational Conference, San Diego, California.

Ten Cate, O. (2016). Entrustment as Assessment: Recognizing the Ability, the Right, and the Duty to Act. *J Grad Med Educ, 8*(2), 261-262. doi:10.4300/JGME-D-16-00097.1

ten Cate, O. (2016). Entrustment Decision-Making in Competency-Based Teaching and Assessment in Health Professions Education. *Medical Science Educator, 26*(S1), 5-7. doi:10.1007/s40670-016-0342-8

Ten Cate, O., Hart, D., Ankel, F., Busari, J., Englander, R., Glasgow, N., . . . International Competency-Based Medical Education, C. (2016). Entrustment Decision Making in Clinical Training. *Acad Med, 91*(2), 191-198. doi:10.1097/ACM.0000000000001044

The College of Family Physicians of Canada. (2020). Certification Examination in Family Medicine. Retrieved from https://www.cfpc.ca/FMExam/

The World Bank. (2020a). GDP growth (annual) - Rwanda. Retrieved from https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2018&locations=RW&start=2000

The World Bank. (2020b). Net official development assistance received (current US$) - Rwanda, Burundi. Retrieved from https://data.worldbank.org/indicator/DT.ODA.ODAT.CD?end=2018&locations=RW-BI&start=1994

The World Bank. (2020c). World Bank Country and Lending Groups. Retrieved from https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

The World Bank. (2020d). The World Bank in Rwanda. Retrieved from https://www.worldbank.org/en/country/rwanda/overview

Thorne, S., Kirkham, S., & O'Flynn-Magee, K. (2004). The Analytic Challenge in Interpretive Description. *International Journal of Qualitative Methods, 34*(1), 1-11. doi:https://doi.org/10.1177/160940690400300101

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care, 19*(6), 349-357.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *European Review of Applied Psychology, 54*(2), 119-135. doi:10.1016/j.erap.2003.12.004

Van Der Vleuten, C. P. (1996). The Assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Sciences Education, 1*(1).

van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: from methods to programmes. *Med Educ, 39*(3), 309-317. doi:10.1111/j.1365-2929.2005.02094.x

van der Vleuten, C. P., Schuwirth, L. W., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Med Teach, 34*(3), 205-214. doi:10.3109/0142159X.2012.652239

van der Vleuten, C. P., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol, 24*(6), 703-719. doi:10.1016/j.bpobgyn.2010.04.001

Vargas, A. L., Boulet, J. R., Errichetti, A., van Zanten, M., Lopez, M. J., & Reta, A. M. (2007). Developing performance-based medical school assessment programs in resource-limited environments. *Med Teach, 29*(2-3), 192-198. doi:10.1080/01421590701316514

Vergel de Dios, J. (2020, March 22, 2020). [Director, CBME Implemntation, Integrated Medical Education].

Whitehead, C. R. (2016). On gunboats and grand pianos: medical education exports and the long shadow of colonialism. *Adv Health Sci Educ Theory Pract, 21*(1), 1-4. doi:10.1007/s10459-015-9660-7

Wikipedia contributors. (2020a, January 11, 2020). Healthcare in Rwanda. Retrieved from https://en.wikipedia.org/w/index.php?title=Healthcare_in_Rwanda&oldid=935209199

Wikipedia contributors. (2020b, March 11, 2020). Rwanda. Retrieved from https://en.wikipedia.org/w/index.php?title=Rwanda&oldid=944983792

Wilbur, K., Hassaballa, N., Mahmood, O. S., & Black, E. K. (2017). Describing student performance: a comparison among clinical preceptors across cultural contexts. *Med Educ, 51*(4), 411-422. doi:10.1111/medu.13223

World Federation for Medical Education. (2017). WFME Global Standards for Quality Improvement: Postgraduate Medical Education. *Postgraduate Medical Education*. Retrieved from https://wfme.org/standards/pgme/

World Health Organization. (2009). *World Health Statistics*. World Health Organizatoin Retrieved from https://www.who.int/whosis/whostat/2009/en/.

World Health Organization. (2020). Key Country Indicators. Retrieved from https://apps.who.int/gho/data/node.cco.ki-RWA?lang=en

Xiao, H., Xian, L., Yu, X., & Wang, J. (2007). Medical curriculum reform in Sun Yat-sen University: implications from the results of GMER evaluation in China. *Med Teach, 29*(7), 706-710. doi:10.1080/01421590701713579

Appendix A – Examiners' Focus Group Interview Guide

Question 1:

- What do you think of the OSCE results of the interns?
- Do you trust the results i.e. think they accurately reflect the performance of the interns on the OSCE?

Question 2:

- There was a variety of opinions on whether to potentially use an OSCE as part of RMDC physician licensing requirements. What is your perspective on this and why?

Question 3:

- Entrustable professional activities (EPAs) are "Important routine care behaviours and activities of a physician that are able to be judged as 'entrustable', which is defined as readiness to safely perform the activity without supervision". You were asked to rate interns performance on following scale as entrustable (or higher) if you felt you would be comfortable for the candidate to carry out the EPA for your patients, in your absence

A (1)	B (2)	C (3)	D (4)	E (5)	F (6)
NOT ENTRUSTABLE For independent practice	BORDERLINE NOT ENTRUSTABLE For independent practice	APPROACHES ENTRUSTABLE For independent practice	ENTRUSTABLE For independent practice	PROFICIENT For independent practice	EXPERT For independent practice

- 'Entrustability' requires a judgement that could be seen as subjective compared to other traditional assessment methods i.e. compared to a pre-defined checklist or written test of items where your score is a numeric one based strictly on the number of items that you completed or got right

- What do you think of using an entrustability scale to rate intern physician performance? Does it make sense to you?
- Do you think it is appropriate? Why or why not?

Question 4:

- Individual trainees may take different amounts of time to reach an entrustable level of performance; this means that even though your time-based training is finished (i.e. 1 year internship), you may not have reached entrustability
- If entrustability is held as a required standard for independent practice, this could have consequences including reducing the number of physicians that are 'practice-ready' at a given time and can be deployed to hospitals as independent practitioners
- Particularly in a limited-resource setting such as Rwanda, what do you think about this? What are some potential benefits versus disadvantages for individual doctors, for the medical profession, for the ministry, for the public?

Question 5:

- Who do you think would be sufficiently qualified or experienced to make entrustability assessments for intern physicians?
- Do you think physicians in District Hospitals, who are the usual 'supervisors' of intern physicians, would be suitable for this? Why or why not?
- Do you think they get enough first-hand exposure to interns' actual practice to accurately make this assessment?

Question 6:

- OSCEs have been considered a relatively resource-intensive method of assessment in terms of time, money, space, organization and people required

- However, it is still used in many places of the world at particular points in training i.e. before graduation from medical school, general licensure, specialist licensure
- Is such a resource-intensive assessment for physicians ‘worth it’, particularly in a limited resource setting? Why or why not?

Question 7:

- It has been suggested that physicians in district hospitals should become more invested in intern training and teaching and that they should make time to supervise, teach, support/mentor and see it as their duty
- However, it is also noted that the clinical demands are so high that this is not realistic
- What are your thoughts on this? What are some other barriers that you see to interns receiving quality training at district hospitals?
- What are your ideas on how these issues should or could be addressed?

Appendix B – Examiners' Post-OSCE Questionnaire (paper survey)

Part 1

Regarding the *Interns' OSCE overall*, please state your level of agreement with the following:

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
The OSCE achieved its stated goal as an objective standardized clinical assessment.					
The OSCE was a useful exercise for assessment and/or training of interns.					
The OSCE content was relevant to the future clinical practice of intern doctors.					
I would recommend running an OSCE again for interns in future years.					
I would be willing to participate as an examiner for an OSCE in future years.					
I think performance on the OSCE could be used as part of RMDC physician licensing process and standards in the future.					

Please provide any general comments or suggestions you have about the OSCE:

Please provide suggestions for future OSCE stations that would be relevant to the clinical practice of general medical officers:

Part 2

Your specialty: _____

Your OSCE Station (please mark 'X' in appropriate station):

Neonatal resuscitation ☐	Gastroenteritis with IO access ☐	Severe malaria ☐	Diabetic Ketoacidosis ☐	Lumbar puncture ☐
Pre-eclampsia ☐	Post-partum hemorrhage ☐	VBAC counseling ☐	Uterine incision closure ☐	Post-operative fever ☐

Regarding the *OSCE station that you were an examiner for today*, please provide feedback on the following:

1. Was the content of your station relevant and realistic to clinical practice?

   ☐ Yes ☐ No ☐ Maybe ☐ Not Sure

2. Was the amount of time allotted for your station appropriate?

   ☐ Too much time ☐ Just about right amount of time ☐ Too little time

3. Was the level of difficulty of your station appropriate for intern physicians?

   ☐ Too easy ☐ Appropriate level of difficulty ☐ Too difficult

4. Would you suggest that we use this station again in the future?

   ☐ Yes ☐ No ☐ Maybe ☐ Not Sure

5. Please provide any general comments or suggestions that you have about your station in the space below:

Appendix C – Former Interns' Survey (administered electronically)

TSAM Post-Internship Survey

Start of Block: Introduction

Q5 This is a survey for the study 'Evaluation of Medical Internship in Rwanda'. You are receiving this because you consented to participate in this study at the time that we conducted the Interns' OSCE (in August 2017 or August 2018).

This is your opportunity to provide us with valuable feedback about your internship experience. The goal is for results to be published and shared with key stakeholders (including the Rwanda Ministry of Health and Rwanda Medical and Dental Council) to evaluate and improve the internship program.

As a reminder, your identity will remain anonymous and confidential. Only de-identified group data will be shared. We expect that the survey will take you up to 15 minutes to complete. The survey will remain open for the next 3 weeks to allow you opportunity to complete it.

We highly value your feedback and appreciate you taking the time to complete the survey. Should you have any questions related to the study or the survey, you may contact the Principal Investigator (Dr. Amita Misir, Western University) directly at amita.misir@lhsc.on.ca.

Thankyou for your participation!

End of Block: Introduction

Start of Block: Identification

Q1 During what time period did you ***COMPLETE*** your general medical internship in Rwanda?

- ○ August - November 2017 (1)
- ○ August - November 2018 (2)
- ○ After December 2018 (3)

Q4 What is your sex?

- ○ Male (1)
- ○ Female (2)

Q6 Where did you complete your internship?

- ○ Northern Province (1)
- ○ Southern Province (2)
- ○ Kigali Province (3)
- ○ Other (4)

Q14 At the time that you did your internship, were you a national or a non-national?

- ○ National (1)
- ○ Non-National (2)

End of Block: Identification

Start of Block: Section 1: Hospital Facilities

Q3 The following items are about the facilities provided to you during your internship year. Please answer as accurately as possible.

Q25 My private living accommodations were acceptable (i.e. secure, in good condition and affordable).

- Yes (1)
- No (2)

Q7 When I was on night duty, there was a clean, private room with a bed to rest that could be secured with a lock and that was available to me within 10 minutes of the hospital.

- Yes, most of the time (1)
- Sometimes (2)
- No, not usually (3)

Q26 The hospital had adequate equipment, services and facilities available to carry out my duties.

- Yes, very adequate (1)
- Fairly adequate but some availability issues (2)
- Somewhat inadequate (3)
- Extremely inadequate (4)

Q8 The hospital had reliable wireless network access.

- Yes, most of the time (1)
- Sometimes (2)
- No, not usually (3)

End of Block: Section 1: Hospital Facilities

Start of Block: Section 2: Internship Program Governance

Q27 The below items are about governance during the internship year that you completed. Please answer the following items as accurately as possible.

Q9 I received the Medical Interns' Handbook (which includes the medical intern logbook) from the Ministry of Health at the start of my internship.

- Yes (1)
- No (2)
- Not Sure (3)

Q10 I was clearly oriented on my training objectives and scope of clinical work throughout my internship.

- Yes, agree this was done adequately (1)
- Somewhat agree, but could be done better (2)
- Do not feel this was done adequately (3)
- Barely done or not done at all (4)

Q28 I received adequate orientation to my hospital(s) and to each department during my internship.

- Yes, agree this was done consistently (1)
- This was done but inconsistently (2)
- This was not done adequately (3)
- This was barely done or not done at all (4)

Q29 I felt that the hospital(s) where I worked were clearly aware of my role and my limitations as an intern.

- Yes (1)
- No (2)
- Unsure (3)

Q30 I knew the internship co-ordinator at my hospital and felt comfortable accessing him/her at any time.

- Yes (1)
- No (2)
- Unsure (3)

Q31 I had adequate health insurance coverage.

- Yes (1)
- No (2)
- Unsure (3)

Q32 I received regular and timely salary/remuneration.

- ○ Yes, always (1)
- ○ Usually, most of the time (2)
- ○ Some of the time (3)
- ○ Rarely or never (4)

End of Block: Section 2: Internship Program Governance

Start of Block: Section 3: Clinical Exposure

Q11 **The below items are about the clinical exposure related to your internship year.** Please answer the following as accurately possible.

	Strongly Disagree (1)	**Disagree** (2)	**Somewhat disagree** (3)	**Neither agree nor disagree** (4)	**Somewhat agree** (5)	**Agree** (6)	**Strongly agree** (7)
My medical school training adequately prepared me for internship. (1)	○	○	○	○	○	○	○
The level of responsibility expected of me was appropriate for my level of experience/expertise at all times during my internship. (2)	○	○	○	○	○	○	○
I was exposed to enough *variety and volume* of patients to prepare me to be a competent general medical officer by the end of my internship year. (3)	○	○	○	○	○	○	○
The more senior physicians (including my direct supervisors) at my hospital took the initiative, time and interest to teach/guide me to become a better doctor. (4)	○	○	○	○	○	○	○

My internship training adequately prepared me for independent practice as a general medical officer in Rwanda. (5)	○	○	○	○	○	○	○

End of Block: Section 3: Clinical Exposure

Start of Block: Section 4: Clinical Supervision/Support

Q13 The below items are about the clinical supervision and support received during your internship year. Please answer the following as honestly as possible.

	Strongly Disagree (1)	**Disagree** (2)	**Somewhat Disagree** (3)	**Neither Agree nor Disagree** (4)	**Somewhat Agree** (5)	**Agree** (6)	**Strongly Agree** (7)
My clinical supervisors were often present and readily available for case review and/or direct supervision. (1)	○	○	○	○	○	○	○
At the start of internship, a more senior physician was assigned to do the night shift with me so that I could learn from him/her and so he/she could ensure that I was managing patients safely. (2)	○	○	○	○	○	○	○
If I needed help, I always had someone to call (day or night). (3)	○	○	○	○	○	○	○

If I had a difficult case, I was encouraged to call someone for help (day or night). (4)	○	○	○	○	○	○	○
If I asked someone at my hospital for help, it was readily given in a supportive way (day or night). (5)	○	○	○	○	○	○	○
The senior physicians at my hospital had sufficient experience and competence to provide safe supervision for me. (6)	○	○	○	○	○	○	○

There were times that I felt that a patient may have been put at increased risk of disability or death due to my relative lack of experience and/or because I did not have appropriate supervision . (7)	○	○	○	○	○	○	○

End of Block: Section 4: Clinical Supervision/Support

Start of Block: Section 5: In-service education

Q15 Did you attend the Internship Refresher Course on Maternal, Newborn and Child Health (MNCH) in Musanze sponsored by Training, Support and Access Model (TSAM) in either Jan-May 2017 or Jan-May 2018?

- ○ Yes (1)
- ○ No (2)
- ○ Unsure (3)

Display This Question:

If Did you attend the Internship Refresher Course on Maternal, Newborn and Child Health (MNCH) in Mu... = Yes

Q16 If you attended the Internship Refresher Course in MNCH, how effective did you feel it was as a learning experience?

- ○ Extremely effective (1)
- ○ Very effective (2)
- ○ Moderately effective (3)
- ○ Slightly effective (4)
- ○ Not effective at all (5)

End of Block: Section 5: In-service education

Start of Block: Section 6: Feedback and Evaluation

Q17 My supervisors and I used the medical interns' logbook during my internship.

- ○ Yes (1)
- ○ No (2)
- ○ Unsure (3)

Q20 I received formal written evaluations at least every 3 months during my internship.

- ○ Yes (1)
- ○ No (2)
- ○ Unsure (3)

Q21 The quarterly (i.e. every 3 months) written formal evaluations were a useful assessment that helped with my professional development.

- ○ Yes, definitely useful (1)
- ○ Maybe somewhat useful (2)
- ○ No, not useful (3)
- ○ Not applicable as I do not recall receiving written formal evaluations every 3 months (4)

Q18 Did you participate in the 10-station Rwandan Interns' Objective Structured Clinical Examination (OSCE) in August 2017 or August 2018?

- ○ Yes (1)
- ○ No (2)
- ○ Unsure (3)

Q19 The below items are about the feedback and evaluation that you received during your internship year. Please answer the following as honestly as possible.

	Strongly Disagree (1)	**Disagree** (2)	**Somewhat disagree** (3)	**Neither agree nor disagree** (4)	**Somewhat agree** (5)	**Agree** (6)	**Strongly agree** (7)
I received regular and timely on-the-job feedback (written or verbal) on my performance throughout my internship year. (1)	○	○	○	○	○	○	○
The feedback was provided in a manner that was helpful in identifying my strengths and weaknesses. (2)	○	○	○	○	○	○	○
I was supported in trying to improve my weaknesses. (3)	○	○	○	○	○	○	○
The evaluations that I received were fair and reflective of my abilities. (4)	○	○	○	○	○	○	○

The Interns' Logbook was a useful and practical way of continuous evaluation during internship. (5)	○	○	○	○	○	○	○
I felt the Objective Structured Clinical Examination (OSCE) was a **useful learning experience**. (6)	○	○	○	○	○	○	○
I felt the Objective Structured Clinical Examination (OSCE) was a **good method of assessment**. (7)	○	○	○	○	○	○	○
I think that an Objective Structured Clinical Examination (OSCE) exercise should be run for future interns. (8)	○	○	○	○	○	○	○

End of Block: Section 6: Feedback and Evaluation

Start of Block: Section 7: Final Questions

Q22 **Overall as a learning experience, I would rate my internship experience as follows:**

- ○ One of the worst learning experiences I have had (1)
- ○ Very poor learning experience (2)
- ○ Below expectations as a learning experience (3)
- ○ Acceptable learning experience (4)
- ○ Good learning experience (5)
- ○ Very good learning experience (6)
- ○ Exceptionally good as a learning experience (7)

Q23 Please use space below to provide any additional comments you wish to share regarding your internship experience.

_____

_____

_____

_____

_____

End of Block: Section 7: Final Questions

Start of Block: End of Survey