QATAR UNIVERSITY

COLLEGE OF PHARMACY

EVALUATION OF A CUMULATIVE EXIT-FROM-DEGREE OBJECTIVE STRUCTURED CLINICAL EXAMINATION (OSCE) IN A GULF CONTEXT

BY

AHMED H. SOBH

A Thesis submitted
to the Faculty of
The College of
Pharmacy in Partial
Fulfillment of the
Requirements for
the Degree of
Master of Science
in Pharmacy

March 2016

COMMITTEE PAGE

The members of the Committee approve the thesis of Ahmed H. Sobh defended on 1 March 2016.

Dr. Kyle John Wilby

Thesis Supervisor

Dr. Mohamed Izham

Thesis Co-supervisor

Dr. Mohamed Diab

Committee Member

Dr. Zubin Austin

Committee Member

Dr. Feras Qasem Alali

Committee Member

Approved:

Dr. Ayman El-Kadi, Dean, College of Pharmacy

Abstract

This study aimed to evaluate the psychometric properties of the 2nd iteration of an Objective Structured Clinical Examination (OSCE) for graduating pharmacy students in Qatar. A secondary objective of this study was to identify quality improvement opportunities for design, implementation, and evaluation of the OSCE.

The psychometric analyses occurred as follows: Cut score determination using borderline regression method; predictive validity using regression and correlation of select course grades and assessments with OSCE scores, concurrent validity using correlation between other cumulative assessments and OSCE scores, risk of bias using correlation between assessors' analytical and global scoring, content validity using student-feedback forms, and interrater reliability using intra-class correlation coefficients (ICCs), and internal consistency using Cronbach's alpha. Pearson and Spearman correlation statistics were conducted at α level < 0.05. A series of two focus groups and subsequent qualitative content analysis were conducted with key stakeholders to identify strengths, weaknesses, opportunities, and challenges regarding OSCE implementation.

Total cut score for the exam was 55.3%. Overall pass rate was 79.2%. OSCE scores correlated moderate-strongly with course grades of Professional Skills and Integrated Case-based Learning, and formative OSCE assessments. Course grades for medicinal chemistry were not correlated with OSCE scores. OSCE scores were moderately predicted by Professional skills course grades (52.3%) and its formative OSCE assessment (61.2%). Average correlation between analytical and global grades for all assessors was 0.52. A total of 90% of the stations were deemed to reflect practice, according to student perceptions. The average intraclass correlation coefficient for analytical checklists scores, global scores, and total scores were 0.88 (0.71 – 0.95), 0.61 (0.19 – 0.82), and 0.75 (0.45 – 0.88) respectively. Cronbach's alpha of students' performance in global scores across stations was 0.87, and 0.93 in terms of total scores. Focus groups confirmed content validity as a weakness yet spoke to training and assessment techniques as both strengths and areas for improvement.

In sum, the 2nd iteration of a cumulative OSCE for graduating pharmacy students in Qatar was deemed valid and reliable, however refinements can be implemented in future iterations to further improve the exam as a high stakes assessment.

Table of Contents

List of Tables ... x

List of Figures ... xi

Abbreviations ... xii

Acknowledgments ... xiii

Chapter 1: Introduction ... 1

1.1 Assessments ... 1

1.2 Types of assessments for learners ... 1

1.3 Performance-based assessments ... 3

1.3.1 Performance-based assessments in pharmacy ... 5

1.4 Objective Structured Clinical Examination (OSCE) ... 10

1.5 Pharmacy and OSCE in the Middle East ... 12

1.6 Brief Introduction about Qatar ... 13

1.7 Problem Statement ... 13

1.8 Adaptation of OSCE in non-Western Cultures ... 16

1.9 Rationale and Research Question: ... 19

1.10 Hypotheses: ... 19

1.11 Objectives: ... 20

1.12 Study significance ... 20

Chapter 2: OSCE Psychometric Evaluation ... 22

2.1 Introduction ... 22

2.1.4 Objectives: 24

2.2 Methodology 25

2.2.1 Research design 25

2.2.2 Methods 26

2.2.2.1 Population 26

2.2.2.2 Variables 26

2.2.2.3 Psychometric analysis 27

2.2.2.3.1 Standard setting: 27

2.2.2.3.2 Normality distribution: 28

2.2.2.3.3 Criterion validity 29

2.2.2.3.3.1 Predictive validity 29

2.2.2.3.3.2 Concurrent validity 30

2.2.2.3.4 Internal validity 31

2.2.2.3.5 Content validity 32

2.2.2.3.6 Inter-rater reliability 33

2.2.2.3.7 Internal consistency 34

2.2.2.3.8 Analytical checklists' items revision 34

2.3 Results 35

2.3.1 Standard setting 35

2.3.2 Normality distribution 38

2.3.3 Criterion validity 42

2.3.3.1 Predictive validity 43

2.3.3.2 Concurrent validity 43

2.3.4 Internal validity 44

2.3.5 Content validity 45

2.3.6 Inter-rater reliability 47

2.3.7 Internal consistency 48

2.3.8 Analytical checklist’s item revision 49

2.4 Summary 50

Chapter 3: Critical Evaluation of the 2015 OSCE: A Qualitative Analysis 52

3.1 Introduction 52

3.1.2 Research Question 52

3.1.3 Objectives 53

3.2 Methodology 53

3.2.1 Study design 53

3.2.1.1 Qualitative analysis 53

3.2.2 Sample selection 53

3.2.2.1 Purposeful sampling 53

3.2.2.2 Sample criteria 54

3.2.2.3 Selection process 54

3.2.3 Data collection 54

3.2.3.1 Purpose of choosing a specific design, focus groups method 54

3.2.3.2 Focus groups 55

3.2.3.3 Focus group process 55

3.2.4 Data analysis 57

3.2.4.1 Transcription 57

3.2.4.2 Content analysis 58

3.2.5 Validity and reliability 58

3.2.6 Researcher bias 59

3.2.7 Ethical approval 59

3.3 Results 59

3.3.1 The process, focus groups and transcription 59

3.3.2 Main themes 60

3.3.3 Coding 60

3.3.4 Content analysis 64

3.3.4.1 Strengths 64

3.3.4.2 Weaknesses 66

3.3.4.3 Opportunities 73

3.3.4.4 Challenges 75

3.4 Summary 78

Chapter 4: Discussion, Limitations, Conclusion, and Future Studies 79

4.1 Discussion: 79

4.2 Limitations 93

4.3 Future studies 93

4.4 Conclusion... 94

References 95

Appendix A: A Template of analytical checklist instrument used in evaluation of students' performance in an OSCE station 112

Appendix B: A template of global scoring instrument used in evaluation of students' performance in an OSCE station 113

Appendix C: Table showing the blue print of the OSCE stations describing their topics, focus, and complexity. 114

Appendix D: Table describes internal consistency... 115

List of Tables

Table1: Borderline regression method showing cut scores for analytical checklist and total station grade, and the passing rate using. 38

Table 2: Shapiro Wilk’s test assessing distribution of candidates’ performance. 39

Table 3: Z scores of skewness and kurtosis assessing distribution of candidates’ performance. 40

Table 4: 2-Tailed Pearson correlation statistics and regression analysis of OSCE grades versus prior undergraduate courses’ grades to assess predictive validity. 43

Table 5: 2-Tailed Pearson correlation statistics of OSCE grades versus different assessments’ grades to assess concurrent validity. 44

Table 6: Internal validity assessed through Spearman correlations between the analytical checklist scores and global scores of each assessor were conducted to assess the risk of bias; average of all correlations was calculated. (*: significant at alpha level 0.05; **: significant at alpha level of 0.01) 45

Table 7: Scores based on (Yes/No) answers of 18 students participated in the mini questionnaire. Average of stations’ satisfaction is calculated. 47

Table 8: Inter-rater reliability using single-measure two-way random ICC, and Pearson’s and Spearman’s correlation of raters scoring in individual stations for 2015’s cycle. Significance are showed on this basis: * when p < 0.05, ** when p < 0.01 48

Table 9: Internal consistency of students’ performance in OSCE exam in terms of total and global scores using Cronbach’s alpha, with sub-analysis of undergraduate students and pharmD students groups. 49

Table 10: Analytical checklists’ items of the lowest score in the whole OSCE exam 50

Table 11: Themes and codes generated from the content analysis of the focus groups. 60

List of Figures

Figure 1: Miller’s Pyramid for learning outcomes (12) 4

Figure 2: Adapted Miller’s Pyramid for learning outcomes (13) 6

Figure 3: Schematic diagram of the thesis research design 26

Figure 4: Scatter plots of the checklist score versus the global score for the ten stations in the OSCE with 26 candidates. Each panel presents the linear regression of checklist score versus global score, the cut value for the global score (equal to 30, vertical broken line), and the corresponding cut value for the checklist score (horizontal broken line) according to the borderline regression method. 37

Figure 5: Histograms showing distribution of grades in different assessments and courses. 42

Abbreviations

OSCE, Objective Structured Clinical Examination

CPH-QU, College of Pharmacy, Qatar University

IRR, Interrater Reliability

ICC, Intra-class Correlation Coefficient

SWOC, Strengths, Weaknesses, Opportunities, and Challenges

SP, Standardized Patients

PharmD, Doctor of Pharmacy

SPSS, Statistical Package for the Social Sciences

SWOT, Strengths, Weaknesses, Opportunities, and Threats

Acknowledgments

Words cannot describe how much I am thankful and grateful to be gifted with such a supervisor and mentor, Dr. Kyle John Wilby. I owe him lot for accepting me as his MSc. Student. I enjoyed our one year and a half of research together. He guided me a lot through my thesis and made it as smooth as possible. I would like to acknowledge the amount of work we have done together where I explored different types of research areas with him, including the projects not included in the thesis. We accomplished a lot in a very short time. I would like to thank him for his continuous motivating and encouraging words, and for believing in me from the very beginning. He is such an amazing supervisor that I wish to have someone like him in my PhD soon. Without him, I would not be here today.

I would like to thank my committee members. First, I would like to thank Dr. Zubin Austin for his time, his support, and his valuable comments and guidance in my projects. I would also like to thank Dr. Feras Alali for believing in me and for his enormous effort to make it possible for me to continue my work in the discipline that I am passionate about. I would like as well to thank Dr. Mohamed Izham and Dr. Mohamed Diab for their time and help in my thesis. I would like to thank Dr. Shane Pawluk for helping me in the moderation of the focus groups during my project.

Special thanks from my heart goes to Dr. Sherief Khalifa and Dr. Ashraf Khalil for the mentorship and support they provided me with, and for the continuous advice they gave me during my whole MSc. Degree, which helped me to improve and to overcome many challenges I faced during my two and a half academic years. I am so grateful for

Dr. Nader Kheir because he was always supportive and kind. I am also thankful for him that he devoted some of his precious time to help me enhance my communication and presentation skills.

I would like to thank all the professors and staff who I learnt from them what benefits my research or my personal skills. I would like as well to thank my colleagues in the masters who encouraged me and provided me with advice and support, Youmna Hassouna, Arwa Hammuda, Ayat Hammad, Ahmad Al Kaddour, Amani Zidan, and Yara Arafat.

I would like to acknowledge the support provided by Qatar University and the College of Pharmacy where I was given an assistantship and I was allowed to attend conferences in Qatar.

Last but not least, deeply from my heart, I am so thankful and very grateful for my parents who always believe in me, who always provide me with all kinds of love and support; without them, I would not be here today. I would also like to thank my close friends, beloved ones, and extended family in Qatar who were very supportive and who made the duration of the MSc. degree move faster and easier. Thank you all for always being there for me.

Dedication

To my loving family:

Ms. Nashwa Abdallah

Mr. Hesham Sobh

Frass, Noha, and Hajer

Chapter 1: Introduction

1.1 Assessments

Development of higher education programming and standards is a key contributor to helping Qatar achieve its national vision. As such there are two main components that are relied upon within higher education. The process of transformation of knowledge and/ or skills from the instructor to the learner is the first component, which is followed by the other component, assessment. Assessment, or evaluation, is a process of collecting data by instructors that gives a description or indication of teaching performance plus the degree of knowledge and/or skills acquired by the learner (1). Assessment contributes greatly to the student's learning process and experience in higher education. Specifically, it should be measuring students' understanding and also broadening areas of their critical thinking and creativity (2). For these reasons, assessments must be carefully crafted to reflect differing student learning styles and also expectations for eventual application of knowledge learned (2). Therefore, assessment types are guided by both instructor's judgments, as well as discipline-specific achievement standards that have become international norms.

1.2 Types of assessments for learners

In higher education, there are three common types of assessments, which instructors can use in the process of learning and evaluation. These include formative, interim, and summative assessments.

1) Formative assessment

Formative assessment depends on a mutual feedback process between the instructors and their students to improve learning and fill gaps in knowledge (3-5). This kind of assessment could be repeated several times during a curriculum (3), where students receive feedback from their instructors about the level of their skills or knowledge and then work towards addressing their points of weaknesses. Instructors also receive feedback about the degree of understanding of their students so they can refine their curriculum or teaching methods to enhance effectiveness (3-5). Formative assessments can consist of a variety of activities, including a class discussion, assignment, homework exercise, informal presentation, internship observation, among others. (4). Ideally, the formative assessment results should remain separate from formal grading that links to summative assessment grades (5).

2) Interim assessment

Interim assessment is more longitudinal in nature, as compared to formative assessment (3, 6). Interim assessments are typically conducted at frequent intervals, if possible (3, 6). The main role of this type of assessment is to continually check students' progress through a feedback system (3, 6). Interim assessments happen at a mid-stage between formative and summative assessments (6). They have similar forms like traditional assessments such as multiple choice and "constructed-response questions" (3) and they can be used to compare between different students (6).

3) Summative assessment

Summative type of assessments are conducted at the end of courses, curriculums, or programs after formative or interim assessments (3, 4). At this stage, their purpose becomes more evaluative than diagnostic, where the primary focus is on evaluating students' performances without the feedback mechanism related to formative and interim assessments (3, 4). For summative assessments, instructors concentrate more on major concepts rather than fine details (3). Summative assessments are employed in various forms, including traditional knowledge-based exams (multiple choice questions) and other emerging assessment methods, such as performance based assessment (3-5).

1.3 Performance-based assessments

Performance based assessment, also known as a type of authentic assessment (7), is assessment that requires students to demonstrate achievements in learning by performing tasks of real practice (8, 9). In such kind of assessments, the purpose is to simulate real life situations and integrate "higher-order thinking skills" necessary for success in real life situations (8-10). These higher order thinking skills allow students to integrate knowledge and skills to demonstrate effective performance on simulated real-life tasks (11). Therefore, performance-based assessments are very important in health professional training, where students are trained to be competent in various professional roles.

Performance-based assessment can be related to educational theory. Miller's Learning Pyramid has been purported as a model for learner progression in clinical contexts. It has four levels of learning outcomes which constructs the framework for clinical assessments (figure 1) (12). The lowest level requires the learner to "know" the basic knowledge that should be used in real life situations, which can be assessed using simple objective measures as multiple-choice, true-false and short answers questions. The second level requires the learner to "know how" to use their knowledge in real cases. This level can be assessed usually by case-based exams and problems solving questions. The next level requires the learners to "show how" they can use this knowledge and apply it in controlled simulations to real life. Finally, the highest level is based on what the learner "does" in real life situations (13).

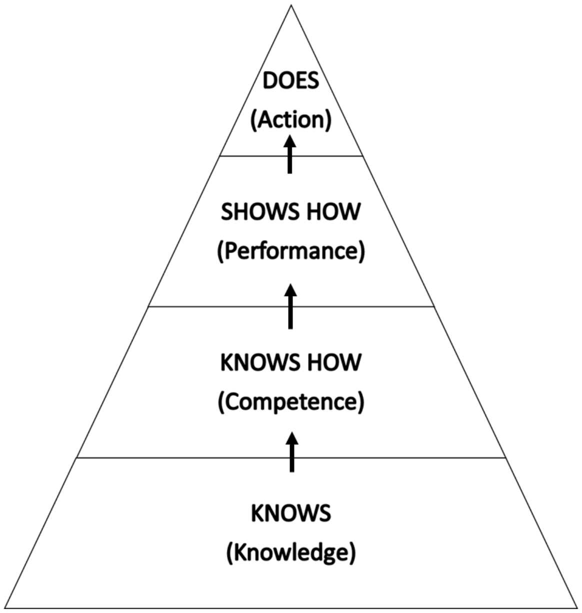


Figure 1: Miller's Pyramid for learning outcomes (12)

The two highest levels of the learning pyramid require the learner to complete multidimensional tasks that rely on integrating knowledge with critical thinking and communication skills. At these two higher levels of learning outcomes, performance-based assessment is essential for assessment of these "higher-order thinking skills". Performance-based assessments allow for simulation of real-life scenarios, which represent the 'shows how' component of the learning pyramid. In a classroom-based setting, this type of assessment is typically the most complex available. Assessments that require students to ‘do’ are primarily limited to clinical internships and experiential training settings.

The reason why performance-based assessments can be considered authentic assessments because they evaluate learners’ performances on “worthy intellectual tasks”, they require learners to perform effectively using the knowledge they learnt, they require these learners to provide comprehensive defensible answers, and they involve tasks that help learners to be prepared for real life situations (14). Such authentic assessments demand a lot of effort and time to judge performances of learners. They require the instructors to decide the type of skills that the students should perform well (7). These skills can be assessed in a number of contexts, such as a simulation, an observation-based rating, a presentation, or a research project (8-10).

1.3.1 Performance-based assessments in pharmacy

Patient-centered care is now standard practice for pharmacy clinicians worldwide. Upon graduation, the pharmacist is expected to perform multiple tasks of increasing complexity beyond just possessing knowledge (13). Pharmacists must be able to integrate knowledge with clinical skills, including provision of patient care, communication, collaboration, advocacy, management, scholarly activity and professionalism (15). Upon graduation, pharmacists must be able to “show how” they integrate knowledge and skills to accomplish the competencies listed above. As such, this can be demonstrated using tasks of real pharmacy practice, according to the adapted Miller’s Learning Pyramid in figure 2 (13).

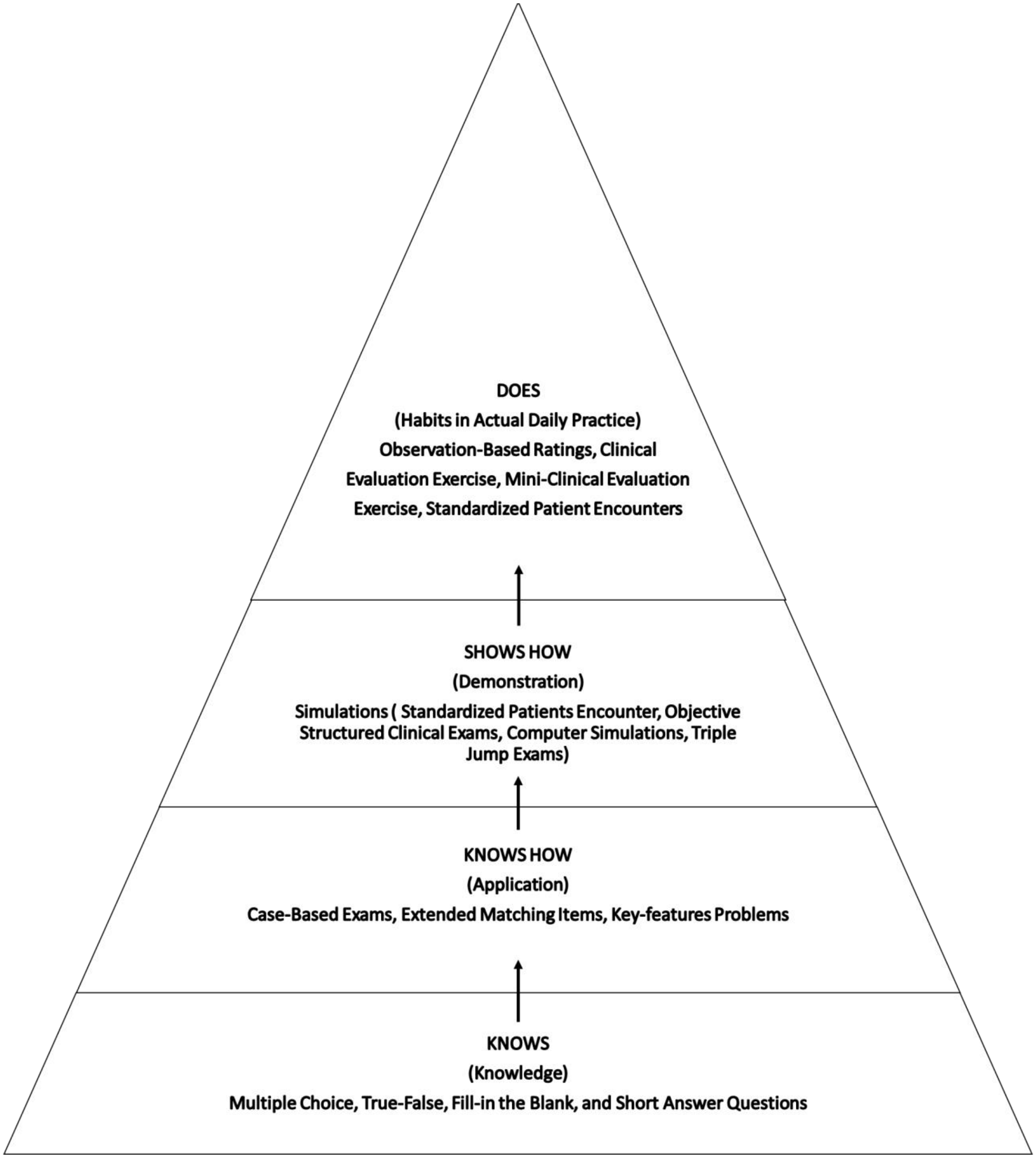

DOES
(Habits in Actual Daily Practice)
Observation-Based Ratings, Clinical Evaluation Exercise, Mini-Clinical Evaluation Exercise, Standardized Patient Encounters
SHOWS HOW
(Demonstration)
Simulations (Standardized Patients Encounter, Objective Structured Clinical Exams, Computer Simulations, Triple Jump Exams)
KNOWS HOW
(Application)
Case-Based Exams, Extended Matching Items, Key-features Problems
KNOWS
(Knowledge)
Multiple Choice, True-False, Fill-in the Blank, and Short Answer Questions

Figure 2: Adapted Miller's Pyramid for learning outcomes (13)

Undergraduate pharmacy students should be well trained on performing real life tasks, in order to have competent pharmacists who excel within patient-care systems. In other words, performance-based assessments should be integrated in the undergraduate pharmacy programs as a part of the learning process, where they will target gaps in the students' knowledge and skills (8, 9, 13). In undergraduate or graduate pharmacy programs, typically, performance-based assessments share the features below (13):

1) The students are required to accomplish a practice-related task
2) The students are familiar from the beginning of the learning process with the criteria and the standards they will be evaluated on during the learning period
3) Assessors directly observe students' performances
4) There could be more than one right answer to the situation or problem
5) Students are assessed on demonstrating their "reasoning skills", where they are required to justify their choice during tasks

Performance-based assessments, as given above, allow students to demonstrate skills and behaviors reflecting real practice. Pharmacy students can be assessed on what they actually do in real practice or what they would do if they were to be in real life situations. If the aim is to evaluate what they actually "DO" in real life practice, then students should be assessed on tasks they perform in real practice (12, 13), and there are different ways, as mentioned in the literature, to assess their performance. For example, they could be observed in a real setting (16, 17), they could be evaluated on accomplished work (13), or even they could be assessed on conducting a research project (18). However, if the aim for a pharmacy student is to "show how" they would perform in a real pharmacy setting, then the kind of assessment mainly depend on simulation (12). They could be assessed in simulations using patients (12), actors (19, 20), or computers (21, 22).

Performance-based assessments gained their importance in pharmacy for the effect they have on improving programs and curriculums (13, 23). However, many challenges have been identified from repeated use:

1) Scoring is a challenge (10, 13) because performance-based tasks usually have more than one correct answer (13, 23) and demonstrate multiple skills; therefore, they require quality judgment by experienced assessors. Assessors who observe the same task could score it differently (13).
2) Time is exceptionally important challenge since preparing, conducting and grading such assessments requires a lot of time by faculty members and assessors. It sometimes requires much time from students in both preparation and task accomplishment (13).
3) Sampling (or determination of which skills or content areas to be assessed) is another challenge especially if time is limited, which means students might only be assessed on fewer tasks covering less skills. As well, some tasks require students to work for days or even weeks (research projects, writing assignments) an therefore are inappropriate to assess in performance-based contexts (13).
4) Resources (human and non-human) are a major barrier in implementation of performance-based assessments, as demands on personnel are high and financial resources are usually required for equipment and/or paying of personnel (24).

In medicine and pharmacy, performance-based assessments have different forms or methods and they can be classified according to their aim. If instructors aim to assess what students actually “DO” in real settings, any of the following assessment methods can be used:

1) Observation-based rating (16), a kind of assessments in which the instructors grade students on their habitual performance in real settings, then they provide the students with feedback in order to improve.

2) Standardized patients (SP) based assessments (10, 13), where actors who act as patients and trained well to evaluate performance of trainees in real settings. This method measures what a trainee (student) actually "does" if SPs are actual patients and they visited the students in real settings and rated their performance directly after the visit (13).

3) Mini-clinical evaluation exercise, or mini-CEX, which requires the trainee to obtain focused patient history and conduct physical examination while being monitored. Following that, students show their work to their assessors where are scored (25, 26).

4) Presentations, or as called, case or patient presentation (16, 27, 28), a type of performance-based assessments that shows how a student can communicate well with a patient, obtain data, critically analyze information, effectively present their case, and respond to other health care professionals' questions (16, 27).

There are other methods as well that describe what a student actually "DOES" in a real setting. They include medical chart audits (29), research projects (18), and learning portfolios (30-32); more information about them could be retrieved from their references.

Demonstration of what students 'can do' if in practice is perhaps the greatest application of performance-based assessment. Typically, these assessments are completed using simulation techniques. Specifically, objective structured clinical examinations (OSCEs) are the most wide spread method of assessment that is used by many academic institutions and health authorities worldwide (22, 33-36). Further detailed explanation about this method will follow in the next section.

1.4 Objective Structured Clinical Examination (OSCE)

OSCEs started in 1975 as a collaboration between the departments of medical education and therapeutics in University of Dundee and the Department of Medicine in Glasgow (37). Nowadays OSCEs are widely known performance-based assessment instruments that are extensively used in various fields of health sciences (pharmacy (38), medicine (39), physical therapy (40), radiography (41), nursing (42), dentistry (43), paramedicine (44), veterinary medicine (45), and others.). OSCE aims to assesses the integration of the learner's knowledge with competence in clinical skills of communication, interactive skills, professional and moral judgments, and problem resolution (33).

OSCE is considered an objective assessment because it relies on more objective measures (checklists, rubrics) to evaluate performance of candidates (23). Typically, the exam consists of a set number of stations, where the candidate encounters a problem to be solved or a task to be completed in a predetermined time (33). Some stations require the candidate to interact with "standardized patients" (actors who are trained to play role of a patient with a disease or a problem) or "standardized clients" (actors or healthcare professionals trained to play role of an "allied" health care professional such as a physician). Static stations without interaction can also be used. Assessors are typically inside the station room with the student directly observing and evaluating the interaction or could perhaps be watching through video software (33).

For educational purposes, OSCEs are used in a formative manner with students or other personnel volunteering to be actors (23), or it can follow the typical structure of OSCE with actors being hired to play roles (46). The main purpose in the formative OSCE is to provide students with feedback about their strengths and weaknesses in a learning process during their curriculums in order to be competent in their fields in real practice. OSCE could also be used as a summative high-stakes assessment where it affects pass/ fail decisions in undergraduate curriculums (13, 33, 47) or even licensure procedures in some countries (48, 49). Therefore, OSCEs must demonstrate high validity and reliability as an assessment instrument (13).

OSCEs have many advantages as an educational assessment instruments:

1) Standardization though hired patient actors decreases variability from use of real patients in authentic practice-based assessments (50).
2) OSCEs allows for flexibility in starting and stopping at set times to provide formative and immediate feedback to students encountering standardized patients (50).
3) Complexity of cases can be modified by exam constructors to fit educational and evaluation purposes (51).
4) Compared to traditional assessments (e.g. multiple-choice questions), clinical skills can be better assessed and measured against pre-defined competencies (50).
5) OSCEs can be a comprehensive instrument by selecting a diverse sample of competencies to assess (52).

6) OSCEs can be used for both formative and summative assessment. Additionally, feedback can be given for improvement even after completion of a summative OSCE (53).

Despite these advantages, OSCEs do also have limitations. Specifically, simulation can never replace authentic assessment in practice. As well, there are many validity and reliability concerns when using simulated actors (54) and unforeseen alternate yet appropriate case solutions that were not accounted for on assessment instruments (54).

Although OSCE has its advantages and disadvantages, it became one of the most popular performance assessment instrument, and it was seen by many as the best assessment of competence (55, 56). However, for an adapted and implemented OSCE to have good impact on learning in the medical and pharmacy education process in an academic institution, it should have an excellent psychometric properties (measures of validity and reliability) (57). Psychometric properties and psychometric analysis will be discussed in full details in chapter 2.

1.5 Pharmacy and OSCE in the Middle East

Although OSCEs are not new to the Middle East (58-66), only few Middle Eastern schools have adapted and used OSCEs in field of pharmacy as a part of their curriculums (67-69). At the College of Pharmacy, Qatar University, the OSCE was adapted from the Canadian context, which is known for its high standards of reliability and validity. It was successfully constructed, implemented and organized in the college. Nevertheless, psychometric analyses from this pilot OSCE in 2014 resulted in poor reliability and validity (70). This warrants the need for additional research and work to be done in order to improve the assessment's validity and reliability.

1.6 Brief Introduction about Qatar

Qatar is one of the smallest Arabic countries that exist in the gulf region. However, it is considered one of the richest countries in the world (71). The country is characterized by a greatly diverse population and workforce consisting of many nationalities and cultures. Recent statistics show that foreigners represent approximately 65% of the population and approximately 94% of its workforce (72). This variety of cultures is reflected in all sectors of the country, including education and healthcare.

With an aim to be an advanced country by 2030, Qatar is investing in every field including education. Qatar targets reaching world-class levels of educational system that promotes analytical and critical thinking (73).

1.7 Problem Statement

Since the advent of patient-centered care in provision of clinical pharmacy services, it became a must to use highly developed performance-based assessment (e.g. OSCE) in the pharmacy education and learning process to train and evaluate students. In order to use such form of assessments, there are three options: either to adopt, adapt, or develop a performance-based assessment that can fit the purpose or learning outcomes required in the College of Pharmacy at Qatar University. Developing a performance-based assessment from the scratch is not a feasible option because it is an expensive process and it will consume a lot of time. Adoption of a test, or an assessment, that has already been created elsewhere is a valuable option because it can save time and money;

however, adopting a performance-based assessment that was initially developed for Western culture and context and conducting it in a gulf country may introduce unintended bias and influence exam validity. Adoption of instruments or assessment methods into new cultural contexts can result in multiple known biases (74):

1) Construct bias: a type of bias that happens because there is no complete overlap between the construct (norms, behaviors, attitudes, etc.) in different groups. This is especially important as constructs typically differ between cultural settings.
2) Method bias: a general term that consists of instrumental bias and item bias.
    a. Instrumental bias: a type of bias that occurs because of the characteristics of the instrument, or assessment, does not relate to the construct of the new culture, which results in score difference due to ethnic differences.
    b. Administration bias: it is a result of a communication problem between the examiner and the examinee. It commonly happens when the both of them are not sharing the same language or using their mother-tongue language or other communication behaviors.

   When compared to the original culture, differences in students' education background, unfamiliarity with techniques of response, and difference in administration conditions (e.g. class size or recording methods) are all sources of method bias.
3) Differential item functioning: a more specific bias that is not related to the whole test or assessment. It is kind of bias because of specific items that does not relate to the new culture or the items have been poorly translated (74).

Since cross-cultural adoption of a performance-based assessment can result in introduction of bias, adaptation became the best option available; improving the cultural appropriateness of the assessment in order to maximize the benefit and the reduce the sources of bias (75). Adapting an assessment promotes fairness in the new culture and context where it is applied, it supports comparison studies between different cultures, and most importantly, it saves money and time needed to develop a new assessment (76).

A typical process of adaption is summarized below (77):

1) First version of the adapted assessment
2) Reviewing the adapted version by reviewers
3) Post reviewing modification
4) Piloting the assessment on a sample of the target population
5) Field test the assessment
6) Scores standardization
7) Perform validation analyses (psychometrics)
8) Develop the assessment's manual and documents for the users
9) Users training
10) Collection of reaction or satisfaction data from users

The previous mentioned procedures are extremely important if the assessment is used for summative purposes; however, some steps can be skipped if the performance-based assessment is mainly used for formative feedback.

To sum up the problem statement in the context of our research, adoption of OSCE (performance-based assessment) from different context and culture could result in several sources of bias. Adaptation of the OSCE, a gold standard of performance-based assessments, is crucial to reduce bias and increase its appropriateness in the gulf context; however, the process should follow the structured procedure given above. This is a novel approach that has not yet been evaluated within the Gulf and likely greater Middle Eastern region. However, adaption data exist from other regions and these experiences are reviewed below.

1.8 Adaptation of OSCE in non-Western Cultures

OSCE has been adapted in several non-Western (outside of North American, European, and Australian) institutions. From a literature review, it appears institutions adapt and use it differently based on their resources and experience. For instance, in Korea, the OSCE has been used in several medical colleges. The exam was conducted in one day. The majority of colleges used residents and hospital staff to play role of standardized patients due to limited resources. They used a limited number (3 or 4) of stations with shorter duration. Once limitation identified in this setting was student sharing of cases and answers, which may compromise exam validity and could be different from Western contexts. (78). This may further increase resource consumption, as new cases must be developed every cycle. While likely not specific to the Korean context, this factor must be considered for OSCE adaption in any setting.

In the United Arab Emirates, the College of Pharmacy of Ras Al Khaimah had a similar experience. OSCE is a part of their community pharmacy course, a course that focuses on pharmacy-based patient care and factors affecting drug selection. In their experience, they developed a blueprint that matches their own learning outcomes for this course. They went through a process of case writing, role-playing using professors, validation and revalidation using medical college staff. They prepared their students using simulated prescriptions in prior laboratory sessions. They tested their stations using students for final revision. Their pilot OSCE had 20 stations, 5 minutes each, with 16 active stations and 4 rest stations. Three of the stations were interactive (i.e. it includes dealing with SP). In general, students found difficulties with patient-problem identification and resolution. There was a general satisfaction by students about the interactive stations. In this study, the students experienced the pilot OSCE for the first time, which may explain the difficulties they faced. This study lacked inter-rater reliability evaluation, which would give an insight about the validity of this adapted assessment in the United Arab Emirates (69).

In Egypt, the OSCE was adapted for the psychiatric nursing program in Alexandria University. It was the first experience for their undergraduate students. The exam was perceived well by the students, yet was deemed stressful by nearly 75% of the students. It can likely be explained by the fact it was their first experience completing such as assessment. The authors in this study conducted inter-rater reliability for the stations that used simulated patients, which were three. It showed moderate reliability (a range from r= 0.581 to 0.708 using Spearman's correlation). In addition to that, they evaluated internal consistency for all stations; although most of the stations showed acceptable internal consistency, other stations were either questionable (α = 0.607), poor (α = 0.582), or unacceptable (α = 0.29 and 0.331) (79). Generally, this was an acceptable adaptation of this performance-based assessment, but more training and experience will be required to make such an exam meet validity and reliability expectations.

Another example comes from Taiwan at the College of Medicine in Kaohsiung Medical University. The college decided to incorporate the OSCE program in their medical educational system to focus on students' clinical skills, communication and attitudes. Based on the main author's experience in the U.S., the college accepted a proposal of establishing the OSCE program. The authors visited a number of top clinical skills centers in U.K, Australia, and the United States. They had lectures being given by experts about the implementation and use of OSCE and standardized patients programs. Some of their medical students were trained to train standardized patients. They recruited standardized patients through advertisements including students, staff and patients. The OSCE they used was formative in nature; they used one observer and three standardized patients taking turns in each station. They used eight OSCE stations. It was a group assessment, where a group of students were been taking history of patients together. The exam process was deemed successful in general. The majority of students were satisfied with the assessment in general (86%) and the improvement in their clinical skills (83%). The study had some limitations; one of these limitations was the use of residents. Students could guess the medical problem of the resident standardized patients because residents acted in case problems that represented their specialty and were known to students. Some residents had conflict of interest with medical students working with them, which could have biased their evaluation (80). The authors did not assess the reliability and validity of their assessment, which could be explained by the fact that their assessment was meant to be formative; however, validity and reliability of their assessment should be evaluated in future cycles.

1.9 Rationale and Research Question:

In these previous experiences mentioned from different settings, it is obvious that the adaptation process is challenging. Although this type of assessments gains acceptance among students and faculty, it is stressful in nature likely due to the novelty within these differing contexts. Additionally, comprehensive validity and reliability analyses were not completed to provide support of the OSCE adaptation as a high stakes exam. Therefore, there is a gap in knowledge of how OSCE can be successfully adapted into non-traditional contexts as a high stakes exam with acceptable psychometric properties.

The basis of this thesis stems from the adaption of a high stakes cumulative OSCE for graduating pharmacy students in Qatar. This exam was developed according to Canadian standards and piloted in 2014. Upon success of the pilot project, the cumulative OSCE was adopted into the curriculum and a second cycle occurred in 2015. The purpose of this project was to evaluate the OSCE and to generate recommendations regarding successful adaption of OSCE into non-traditional contexts. Our specific research questions were:

- What is the validity and reliability of a 2nd iterative cycle of a cumulative, summative OSCE for graduating pharmacy students in a GCC context?
- How can the OSCE be further refined to improve validity and reliability within the GCC context?

1.10 Hypotheses:

We hypothesized that:

1) The 2nd iterative OSCE cycle will have acceptable psychometric properties

2) The 2^{nd} OSCE cycle will identify further refinements required for improving validity and reliability as a high stakes examination

1.11 Objectives:

The following chapters outline methods and results according to the following specific objectives:

1) To determine cut scores and associated pass rate of the 2015 OSCE
2) To determine the predictive validity of performance on the OSCE using formative course grades and performance-based assessments
3) To determine concurrent validity of performance on the OSCE using summative course grades and summative assessments
4) To determine internal validity of the OSCE (risk of assessors' bias)
5) To determine candidate perceptions regarding exam validity
6) To determine exam reliability in terms of internal consistency, inter-rater reliability (for both analytical and global scoring components)
7) To revise checklist items according to performance levels
8) To critically analyze the OSCE from stakeholder perspectives including candidates, assessors, standardized actors, and exam center staff

1.12 Study significance

This project will serve as the first of its kind to comprehensively analyze a 2^{nd} cyclic iteration of a summative, exit-from-degree, cumulative OSCE in a non-Western context. Methods and results can provide a “gold-standard” approach for colleges adapting or planning to adapt OSCE in non-Western countries on how to evaluate the psychometric properties of OSCEs adapted and how to plan for their improvement. The findings will also be relevant to traditional Western contexts, as the current trends in globalization and immigration are increasing multiculturalism within these settings. Finally, results align with the National Vision of Qatar to establish modernized education and health systems (73). By establishing a valid and reliable performance-based assessment in the country, future initiatives may include consideration for health professional licensure or continual competency assessment.

Chapter 2: OSCE Psychometric Evaluation

Definitions:

- Analytical checklist: It is a checklist instrument based on an objective judgment from evaluators. The checklist consisted of tasks that ideally should be performed by the students in each station in order to manage the case successfully (81). It includes tasks of gathering information, disease or case management, and follow up (Appendix A). Any task completed by the student is checked by the station's assessors.
- Global scale scoring, or global scoring: It is an instrument that has a scale, mostly from 1 to 5, used to evaluate the student's overall performance in his/her station, and their communication and global skills (Appendix 2) (81). It is a subjective measure that depends on the assessor's judgment of the student's performance. The higher the score (5 or close) that the students can get on the global scale depends on how well they perform in their station.

2.1 Introduction

This chapter will answer the first research question, "What is the validity and reliability of a 2nd iterative cycle of a cumulative, summative OSCE for graduating pharmacy students in a GCC context?" Before explaining methodology and results of the psychometric analysis, some background information must be given on the OSCE development process in our setting.

In 2014, CPH piloted the first high stakes cumulative OSCE for pharmacy students in Qatar in collaboration with the Supreme Council of Health in Qatar and consultants from the University of Toronto. The OSCE was implemented in response to accreditation recommendations, as well as an identified need to assess program-learning outcomes in a summative manner. The exam was designed as an exit-from-degree exam that aimed to assess students according to a blueprint (Appendix C) based on AFPC competencies (15) and curricular mapping. Upon the success of the pilot project, the OSCE was adopted into the curriculum and repeated in 2015.

The OSCE development process was the same for both 2014 and 2015 cycles, aside from the standard setting process as described below. Cases (according to the developed blueprint) and standardized actor scripts were developed by groups of 4-6 people consisting of faculty, practice pharmacists, and/or regulators. Each group completed a case template and analytical checklist (answer key). Subsequently, a different group received the developed case for validation. During validation, the second group was expected to role-play the case and identify inaccuracies, need to clarification, and/or need for further details. The validation groups were allowed to change any aspect of the case template or analytical checklist. In 2014, the case was passed to a third group for standard setting using the Angoff method (82), however a different method for standard setting was used in 2015, as described below.

In addition to case development and validation, the exam required recruitment and training of assessors, standardized actors, and exam center staff. For the 2015 cycle, assessors were trained over a 2-hour session using a series of calibration exercises and discussion. Attempts were especially made to standardize assessments using the global assessment, as this measure was deemed suboptimal in 2014. In addition to this training session, all assessors and standardized actors for each station met for 2 hours prior to the exam to receive and learn the case, role-play the case, and discuss action plans according to potential student responses. Standardized actors were recruited from the college pool of actors that regularly contribute to formative OSCE assessments in professional skills courses.

For the 2015 cycle, 10 stations were implemented that required students to interact with standardized actors. At least 2 pharmacist-assessors were present (1 station had 3) inside the station to grade students' analytical and global skills and 2 standardized actors switched off for each station aside from 1 station where 1 standardized actor completed all interactions. Examples of assessment instruments are given in Appendix A and B.

In 2014, data demonstrated poor validity and reliability, largely due to the pilot nature of the first cycle. Specifically, examiners felt the Angoff method of standard setting was not appropriate for the cultural context and resulted in an inflated pass rate. Also, inter-rater reliability between assessors for both analytical and global performance was average at best (Intra-class correlation coefficient: 0.77 and 0.48 respectively). Therefore, a comprehensive psychometric analysis was warranted to further understand validity and reliability of the OSCE in our setting.

2.1.4 Objectives:

This chapter evaluates the psychometric properties of the 2015 (2nd cycle) of the cumulative OSCE at CPH. Specific objectives were the following:

1) To determine cut scores and associated pass rate of the 2015 OSCE

2) To determine the predictive validity of performance on the OSCE using formative course grades and performance-based assessments

3) To determine concurrent validity of performance on the OSCE using summative course grades and summative assessments

4) To determine internal validity of the OSCE (risk of assessors' bias)

5) To determine candidate perceptions regarding exam validity

6) To determine exam reliability in terms of internal consistency, inter-rater reliability (for both analytical and global scoring components)

7) To revise checklist items according to performance levels (risk of assessors' bias)

2.2 Methodology

2.2.1 Research design

In order to measure the objectives of our study, the research had to pass through different steps listed in figure 3 below.

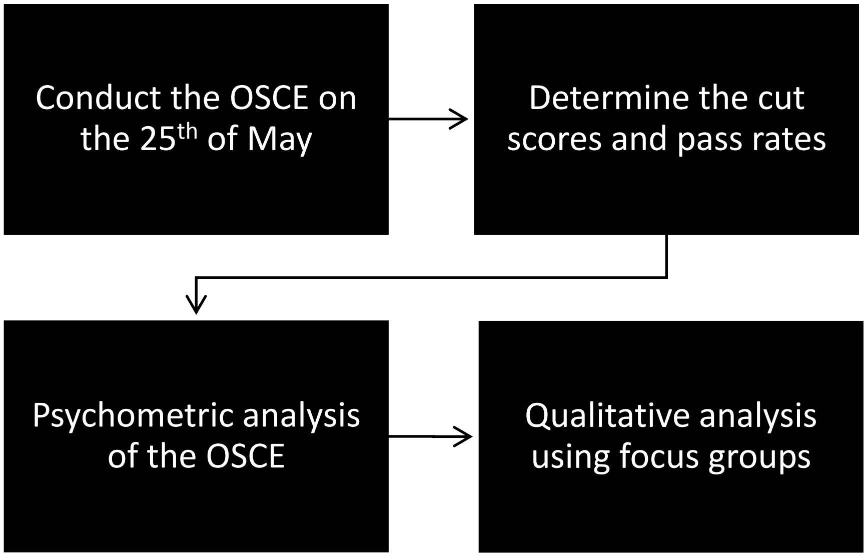

Conduct the OSCE on the 25th of May
Determine the cut scores and pass rates
Psychometric analysis of the OSCE
Qualitative analysis using focus groups

Figure 3: Schematic diagram of the thesis research design

2.2.2 Methods

2.2.2.1 Population

The psychometric analysis was based on student performance and results obtained from exam assessors. The assessors contributed in this exam were a mixture of external assessors selected from different institutes in Qatar such as hospitals and health facilities and internal faculty assessors from CPH-QU. All 43 assessors were pharmacists. Exam candidates were 21 female graduating pharmacy students in their fourth (last) year in CPH-QU and 5 part-time PharmD ((Doctor of Pharmacy) students entering the internship phase of their training) (83). The majority of the candidates were Arab expatriates as Egyptians, Syrians, Lebanese, Palestinian, and Sudanese with one Qatari student.

2.2.2.2 Variables

The following variables must be defined:

1) PHAR201: Medicinal Chemistry (taken in the first professional year of pharmacy)

2) PHAR440, PHAR441: Professional skills courses (taken in the third professional year of pharmacy)

3) SMSA (Structured Multi-Skill Assessment): formative OSCE adapted to fit the outcomes of the curricula in QU (46). It is the performance-based assessment of the professional skills courses, PHAR440 and PHAR441

4) PHAR491, PHAR590: Integrated case-based learning courses, which the students took in their third (PHAR491) and fourth year (PHAR590). The OSCE assessment grades represents 20% of the final grades of PHAR590. There is a cumulative MCQ assessment (also part of the final cumulative assessment) that is a part of the same course, PHAR590.

5) cGPA, or cumulative grade point average: It is the cumulative grade of a pharmacy student during the entire bachelor program. It is a scale up to 4.0.

6) Prometric exam: It is a type MCQ assessment used in Qatar to obtain a pharmacist license. It has different sections, including pharmacology, biopharmaceutics, calculations, and pharmacy practice and clinical pharmacy.

2.2.2.3 Psychometric analysis

Psychometric analysis was done for all stations included in the OSCE. There were a total of 10 active stations that required students to interact with a standardized actor to solve a case. All statistical analyses mentioned below were done using SPSS statistics software version 22 and Microsoft® Excel© version 2013.

2.2.2.3.1 Standard setting:

Cut scores and pass rates were calculated for all OSCE stations using the data of the 26 examinees. Cut scores were calculated using borderline regression method (84), a robust, defendable, and less time-consuming method that uses few resources (85, 86). Using this method, cut scores were determined as following:

1) For each station, using the global score for each student (X-axis) versus the analytical score (Y-axis), scatter plot was generated.
2) Regression line equation was determined from the plots generated.
3) Since 50% of the passing grade was allocated for global scoring and the other 50% for analytical scoring, using 30 out of 50 (3 out of 5) as the pass score of global assessment, we determined the cut score for the analytical checklist through the equations created.
4) Cut scores for the total station grade was determined by adding 30, the pass score of the global rating, to the analytical rating cut score.
5) Pass rates for both analytical rating and total rating per each station were then calculated using student scores on each station.

2.2.2.3.2 Normality distribution:

Normality distribution was determined for all variables to guide choice of analysis (87-89). The variables tested for their normality distribution included the final grades of the courses PHAR201, PHAR440 and its SMSA, PHAR441and its SMSA, PHAR491, and PHAR590, cGPA, Prometric exam grades, the pharmacy practice and clinical pharmacy part of the Prometric exam, the MCQ portion of the final cumulative assessment, and finally the OSCE grades.

Distribution of data were assessed using 3 methods:

1) Shapiro-Wilks test: If the p value of any data is less than the alpha level 0.05, then the null hypothesis is rejected and data are not normally distributed (90, 91). However, if the p value is more than the alpha level 0.05, then we fail to reject the null hypothesis and data are assumed to be normally distributed.
2) Z scores: Z scores were calculated by dividing each of the skewness and kurtosis data of each variable by their standard error values. If the values lie between 1.96 and -1.96, this indicates that data are approximately normally distributed (92).
3) Histograms: Histograms of the different variables were visually inspected to confirm the results of the previous methods. It is the simplest and easiest way to check normality of data. If data resemble bell-shaped curves, it indicates that the data are normally distributed (91).

2.2.2.3.3 Criterion validity

Criterion validity is the degree by which a measure is related to an outcome. It is measured through predictive and concurrent validity, as described below (93).

2.2.2.3.3.1 Predictive validity

Predictive validity is the degree by which a score in a test or a scale could predict a measure of outcome in the future (94). In this study, we attempted to determine the variables that can predict performance of students in the OSCE. We conducted predictive validity as following:

1) We selected the undergraduate courses that share similarities with the skills of communication and critical thinking required to perform well in OSCE. Variables included: 2 professional skills courses (PHAR440 and PHAR441) along with their

SMSAs and 2 integrated case-based learning courses (PHAR491 and PHAR590). The medicinal chemistry course (PHAR201) was selected as a control group. Additionally, student admission rankings upon entrance to the pharmacy program were assessed. This provided a total of 8 variables utilized to measure predictive validity.

2) Pearson correlation was used to identify associations between the grades of students in the variables given above and their grades in the OSCE (95). Alpha level was set at 0.05 for significance.

3) In addition, Spearman correlation was used to determine the effect of student admission ranking to the college on the students' OSCE grades. Alpha level was again set at 0.05 for significance.

4) Any variable significantly correlated using analysis techniques above, a univariate linear regression analysis was done to determine degree of prediction.

2.2.2.3.3.2 Concurrent validity

Concurrent validity measures a degree of association of a particular test or measure with previously established validated test (94). In our study, we determined if there was any degree of association between different types of assessments or scores and OSCE grades. In order to determine concurrent validity, we went through the following steps:

1) The identified variables included: students' cGPA, Prometric grades, the pharmacy practice and clinical pharmacy part of Prometric exam, and the cumulative MCQ portion of the final cumulative assessment.

2) Concurrent validity was measured using Pearson correlation between OSCE grades and each of the previously mentioned variables. Alpha level was set to 0.05 for significance.

2.2.2.3.4 Internal validity

Internal validity is a type of validity that focuses on how well an experiment or a method is done without interference of any other confounding factor or bias (96). It is important to determine if the grades that the students received in the OSCE by assessors was mainly because of their performance in their stations, not because of any other confounding factor or bias. Since each assessor was responsible for evaluating 26 students using 2 instruments, analytical checklist and global score, there was a risk of bias in scoring students in the global scale based on their performance in the analytical checklist.

In order to test the hypothesis that assessors had bias in evaluating students, we completed the following steps:

1) Data of evaluation of the 26 students in terms of analytical and global scoring were collected for each assessor.
2) Spearman correlation statistics was done between both types of scoring for each assessor.
3) Average of all scoring correlation for the 21 assessors was calculated.
4) We assumed that high correlation between the analytical and global scoring for assessors would be an indication of a risk of bias, while moderate or weak correlations would be indication of minimal risk of bias.

2.2.2.3.5 Content validity

Content validity, in the context of our research, is the degree by which elements of an assessment or an instrument can represent variables that fit the exam purpose (97). In other words, how much content in terms of tasks in different cases (elements) in the OSCE (assessment) represents or assesses competencies required upon graduation for eventual integration into practice. Although creating the blueprint and undergoing the case validity process were considered a part of content validity, we further assessed it using the following:

1) For the first method, questionnaire forms were distributed to students one week after completing the OSCE.
2) The questionnaire asked students if they believe the station was reflective of practice (Yes, No, I don't know) and allowed students to provide comments
3) Validation forms were collected, scores were calculated based on their yes or no answers in an Excel sheet, where Yes was considered as 1, No as 0, and "I don't know" as 0.5.
4) Total scores for each station was summed up for each station, divided by the number of responders, and multiplied by 100 to give a percentage of the degree of satisfaction and resemblance to practice in each station from students' perspective.
5) Students' comments were documented, summarized, compared among each other, and analyzed.
6) The second method of assessing content validity was through a qualitative analysis of strength, weaknesses, opportunities, and challenges of the OSCE

conducted in the college using focus groups. This method will be discussed in Chapter 3.

2.2.2.3.6 Inter-rater reliability

Inter-rater reliability or inter-rater agreement is the extent of which two raters or more agree on scoring an outcome measure (91). If two raters were provided with a good instrument of assessment and were provided proper training, ideally, this will result in high inter-rater reliability, which is required to promote validity in assessment. It can be assessed using different methods such (98): joint probability of agreement (99), kappa statistics (100), correlation statistics (101), or intra-class correlation coefficients (ICC) (98). The later one is the most appropriate for the data obtained from the OSCE. In addition, it puts in consideration the differences in raters' evaluation plus their ratings' correlation. In this study, we determined the inter-rater reliability as following:

1) Using SPSS software, ICCs were calculated for both global and analytical rating among assessors in each station using single-measure two-way random intraclass correlation (102), because a fixed sample of raters were used to rate all students in each station.

2) Average ICC of global and analytical rating was calculated for the whole OSCE, followed by overall ICC of the whole exam.

3) As a confirmatory analysis, Pearson correlation was calculated for analytical ratings between assessors (parametric data) and Spearman correlation for global ratings between assessors (non parametric data) to compare with other studies that use Pearson and Spearman correlations.

2.2.2.3.7 Internal consistency

Internal consistency is reliability that focuses on outcome measures for a test, instrument, or large assessment, where it evaluates the degree to which all items in the instrument or assessment evaluate the same core concept (103). Relating this metric to assessment, exam items that assess the same concept (i.e. communication skills) should produce similar scores. In the OSCE, students were assessed on their overall performance and communication skills in every station. Therefore, in our study, we wanted to assess how the exam instruments are consistent in assessing their overall performance in addition to their communication skills for all stations (104). We deemed an acceptable internal consistency measure to be $0.8 > \alpha \geq 0.7$ as shown in appendix D (105, 106). In order to conduct this analysis, we have done the following:

1) In SPSS, we used Cronbach's alpha to measure internal consistency of the 26 students' scores in all 10 station in terms of both global scoring and total scoring.
2) We sub-analyzed the data for the 21 undergraduate students and the 5 graduate students as a sensitivity analysis.

2.2.2.3.8 Analytical checklists' items revision

As described above, analytical checklists were created during the case development and validation processes. Revision of checklist points is necessary to determine how assessed items reflected content and skills taught in the undergraduate program an expected upon graduation. The process for this metric is described below:

1) Overall success on individual checklist items was calculated using Microsoft® Excel v. 2013

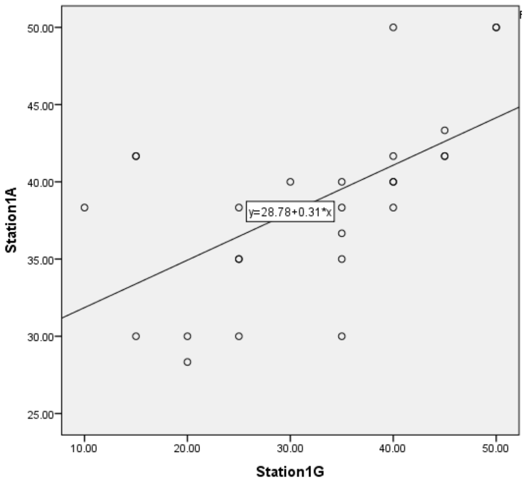
2) Checklist items per station were ranked from highest to lowest achievement

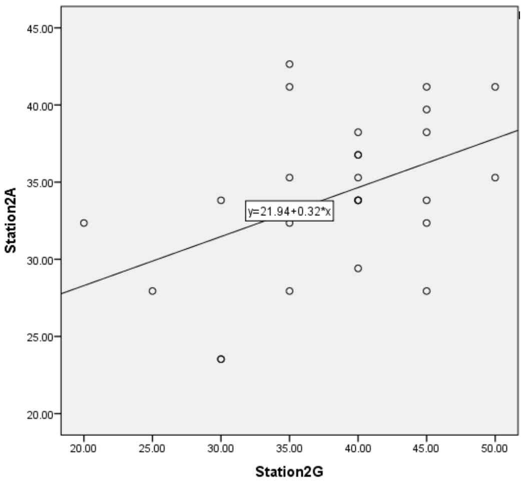
3) Any item that had an achievement rate of 10% or less was further analyzed to determine validity in future cycles or to hypothesize gaps in curriculum that should be addressed to ensure success in future cycles

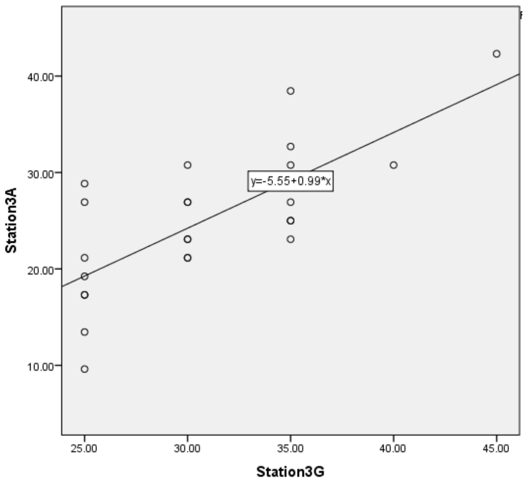
2.3 Results

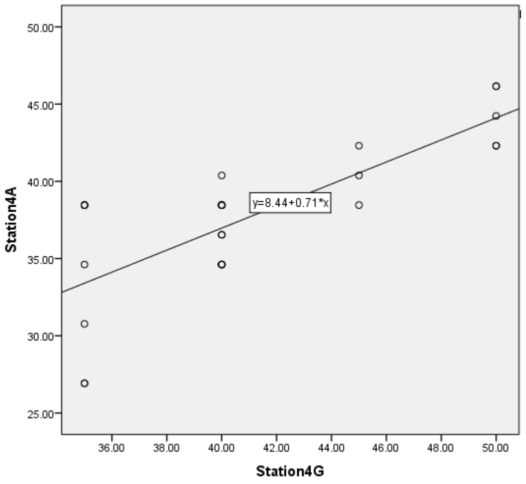
2.3.1 Standard setting

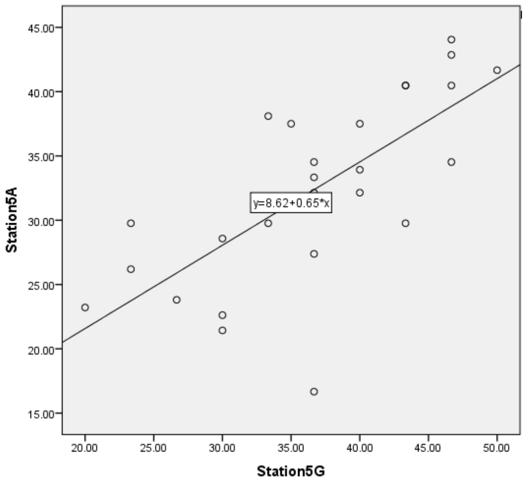
For the ten stations in the OSCE, figure 4 demonstrates the scatter plots of the checklist scores versus the global scores for the 26 students completing the OSCE. Every circle represents the result of a student. Nevertheless, scores of some students are identical and result in overlapping circles in the scatter plot. Every panel in figure 4 shows the linear regression of checklist score versus global score. The equations generated are presented in appendix 6, where checklists cut scores are calculated using 3 out of 5 (30 out of 50) as global scale pass score. In the ten stations, we added 30% of the global cut score to the checklists' cut scores calculated. The cut score of the whole exam was 553.01/ 1000 (55.3%). Data shows that station 1 shared the highest cut score with 68.08%, while station 8 was the lowest with 46.99%. Pass rate calculated per analytical checklist resulted in an average of 70.38%; station 4 showed the highest pass rate per analytical checklist (92.31%) and station 6 showed the lowest with exactly 50%. Average pass rate for the whole exam was 79.23%. Interestingly, all the 26 students passed station 4 (100%) while only 57.69% of them passed station 1.


Station1A
Station1G
y=28.78+0.31*x
50.00
45.00
40.00
35.00
30.00
25.00
10.00
20.00
30.00
40.00
50.00


Station2A
Station2G
y=21.94+0.32*x
45.00
40.00
35.00
30.00
25.00
20.00
20.00
25.00
30.00
35.00
40.00
45.00
50.00


Station3A
Station3G
y=-5.55+0.99*x
40.00
30.00
20.00
10.00
25.00
30.00
35.00
40.00
45.00


Station4A
Station4G
y=8.44+0.71*x
50.00
45.00
40.00
35.00
30.00
25.00
36.00
38.00
40.00
42.00
44.00
46.00
48.00
50.00


Station5A
Station5G
y=8.62+0.65*x
45.00
40.00
35.00
30.00
25.00
20.00
15.00
20.00
25.00
30.00
35.00
40.00
45.00
50.00

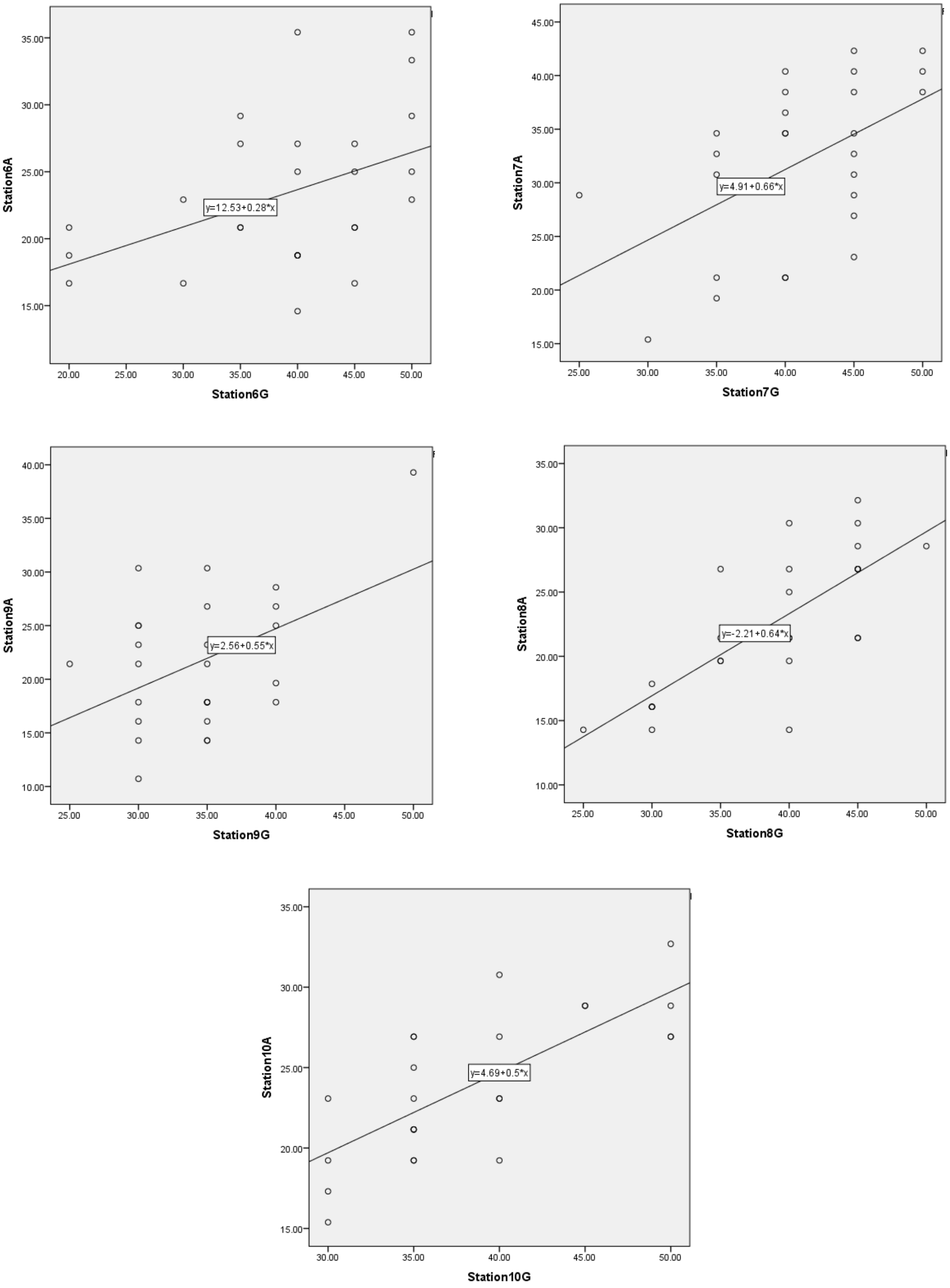


Figure 4: Scatter plots of the checklist score versus the global score for the ten stations in the OSCE with 26 candidates. Each panel presents the linear regression of checklist score versus global score, the cut value for the global score (equal to 30, vertical broken line), and the corresponding cut value for the checklist score (horizontal broken line) according to the borderline regression method**.**

Table1: Borderline regression method showing cut scores for analytical checklist and total station grade, and the passing rate using.

Station	Regression line equation	Analytical checklist cut score	Total cut score	Pass rate per analytical score	Pass rate per total score
Station 1	Y = 28.78+0.31*X	38.08	68.08	65.38	57.69
Station 2	Y = 21.94+0.32*X	31.54	61.54	76.92	84.62
Station 3	Y = -5.55+0.99*X	24.15	54.15	57.69	53.85
Station 4	Y = 8.44+0.71*X	29.74	59.74	92.31	100.00
Station 5	Y = 8.62+0.65*X	28.12	58.12	73.08	73.08
Station 6	Y = 12.53+0.28*X	20.93	50.93	50.00	84.62
Station 7	Y = 4.91+0.66*X	24.71	54.71	76.92	88.46
Station 8	Y = -2.21+0.64*X	16.99	46.99	76.92	80.77
Station 9	Y = 2.56+0.55*X	19.06	49.06	57.69	80.77
Station 10	Y = 4.69+0.5*X	19.69	49.69	76.92	88.46
Average			55.3	70.38	79.23

2.3.2 Normality distribution

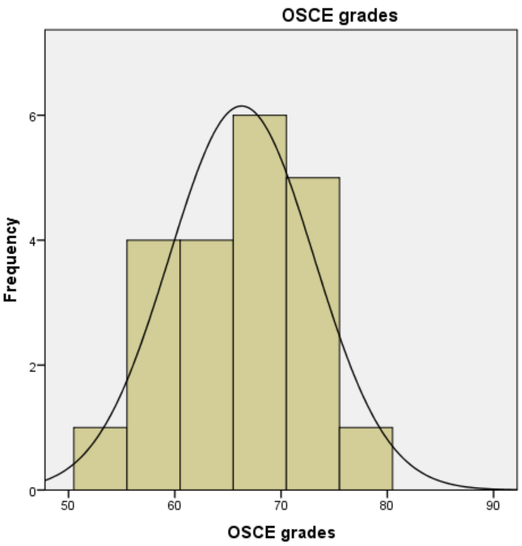
As shown in table 2, using Shapiro-wilk’s method, OSCE grades, cumulative GPA, Prometric grades, professional skills laboratory PHAR441 Total score, PHAR441 SMSA, professional skills laboratory PHAR 440 total, medicinal chemistry, integrated case-based learning courses PH590 and PH491, Prometric pharmacy practice and clinical pharmacy, PHAR590 MCQ, and PHAR440 SMSA showed $p > 0.05$, which indicates that data are normally distributed. These results were confirmed by Z scores, as shown in table 3, where all values lied between -1.96 and 1.96 confirming that all data are approximately normally distributed. Visual inspection through histograms showed normal bell-curve in all data (Figure 5), which confirms all the previous results. This led us to use parametric analysis for the following tests.

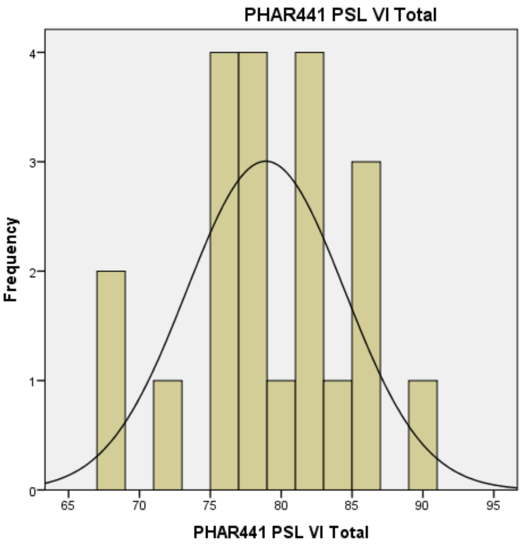
Table 2: Shapiro Wilk's test assessing distribution of candidates' performance.

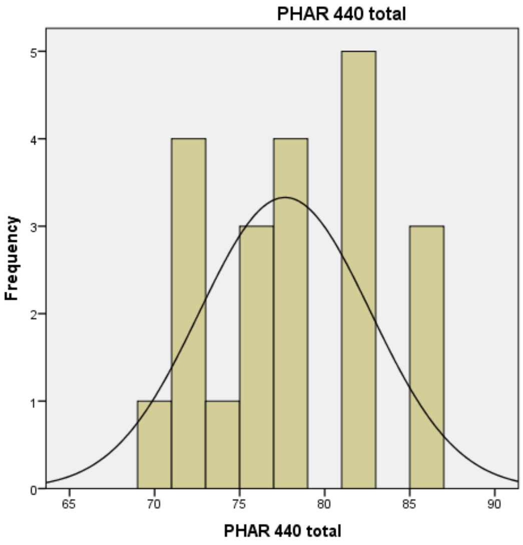
	Gender	Statistic	df	Sig.
OSCE grades	Female	.982	21	.945
cGPA	Female	.967	16	.784
Prometric scores	Female	.955	16	.573
PHAR441 Total	Female	.966	21	.633
PHAR441 SMSA	Female	.970	21	.744
PHAR 440 Total	Female	.939	21	.211
PHAR 201 Medicinal Chemistry Total	Female	.959	25	.404
PH590 ICBL Total	Female	.932	21	.150
PH491 ICBL Total	Female	.955	21	.427
Prometric – Clinical Pharmacy and Pharmacy Practice	Female	.907	16	.103
PHAR590 MCQ	Female	.967	21	.662
PHAR440 SMSA	Female	.953	21	.394

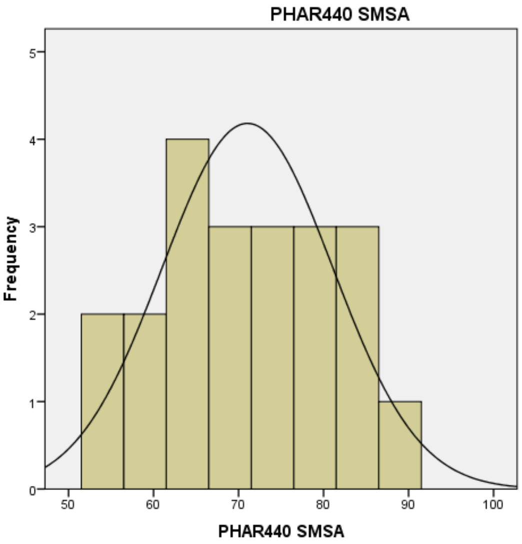
Table 3: Z scores of skewness and kurtosis assessing distribution of candidates' performance.

Course	Skewness	Std. error	Z value	Kurtosis	Std. error	Z value
OSCE grades	-0.021	0.501	-0.04192	-0.408	0.972	-0.41975
cGPA	-0.464	0.564	-0.8227	-0.42	1.091	-0.38497
Prometric scores	-0.685	0.564	-1.21454	0.351	1.091	0.321723
PHAR441 Total	-0.342	0.501	-0.68263	-0.189	0.972	-0.19444
PHAR441 SMSA	-0.024	0.501	-0.0479	-0.949	0.972	-0.97634
PHAR 440 Total	0.193	0.501	0.38523	-1.098	0.972	-1.12963
PHAR 201 Medicinal Chemistry Total	0.428	0.464	0.922414	-0.397	0.902	-0.44013
PH590 ICBL Total	0.199	0.501	0.397206	-1.09	0.972	-1.1214
PH491 ICBL Total	0.007	0.501	0.013972	-0.846	0.972	-0.87037
Prometric – Clinical Pharmacy and Pharmacy Practice	-0.467	0.564	-0.82801	-0.967	1.091	-0.88634
PHAR590 MCQ	0.132	0.501	0.263473	-0.836	0.972	-0.86008
PHAR440 SMSA	0.032	0.501	0.063872	-1.011	0.972	-1.04012


OSCE grades
Frequency
0
2
4
6
50
60
70
80
90
OSCE grades


PHAR441 PSL VI Total
Frequency
0
1
2
3
4
65
70
75
80
85
90
95
PHAR441 PSL VI Total


PHAR 440 total
Frequency
0
1
2
3
4
5
65
70
75
80
85
90
PHAR 440 total


PHAR440 SMSA
Frequency
0
1
2
3
4
5
50
60
70
80
90
100
PHAR440 SMSA

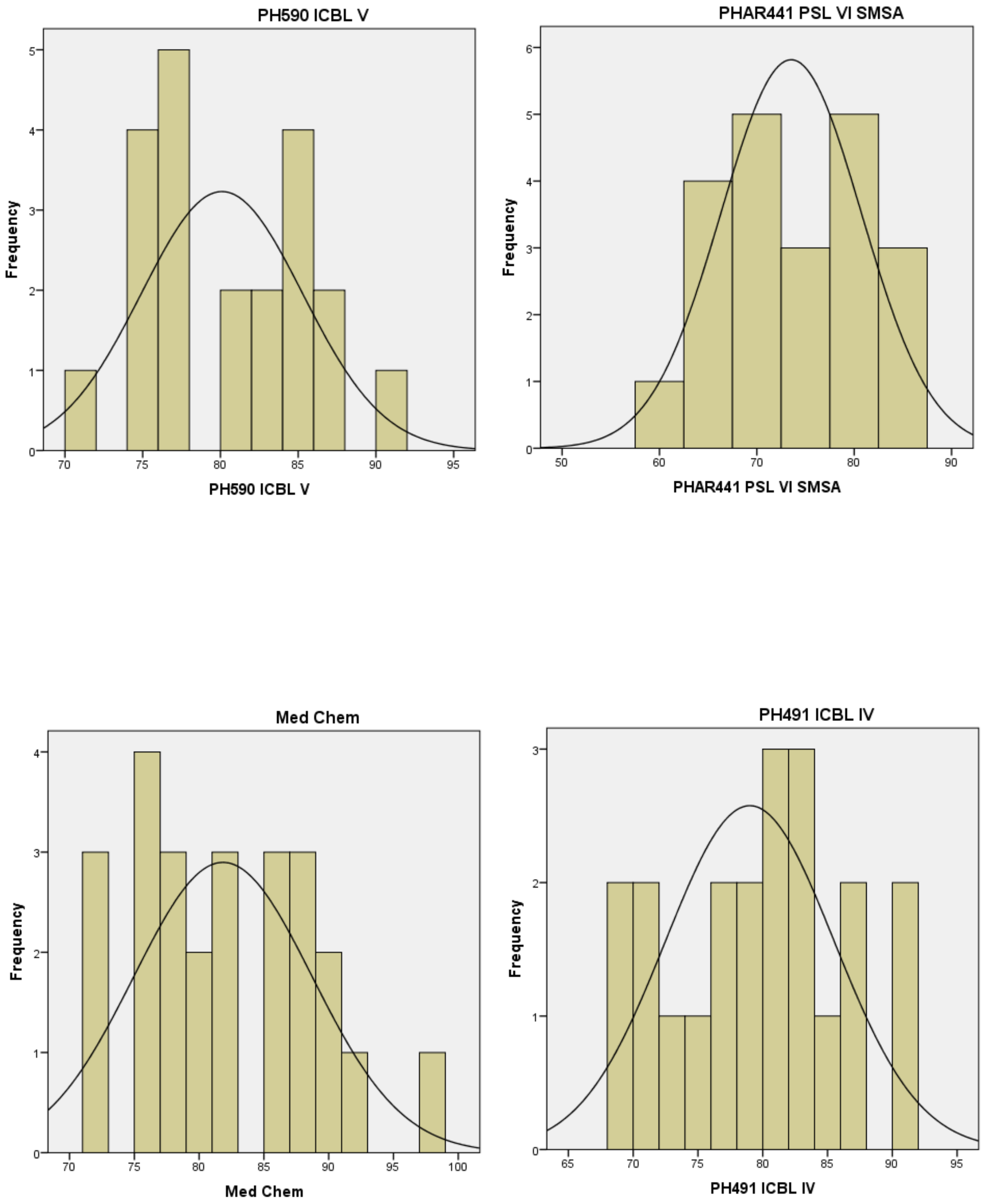


Figure 5: Histograms showing distribution of grades in different assessments and courses.

2.3.3 Criterion validity

2.3.3.1 Predictive validity

For the variables determined to predict OSCE performance, PHAR 441, PHAR 441 SMSA, PHAR 440, and PHAR590 showed significant moderate to good positive correlation, as shown in table 4, with one of the formative OSCEs, SMSA of PHAR440, showed significant strong positive correlation ($r = 0.78$, $p < 0.01$). Only one course showed moderate non-significant positive correlation, which is one of the integrated case-based learning courses, PHAR491. The medicinal chemistry course's grades, that acted as a control, showed weak positive correlation ($r = 0.34$, $p > 0.05$) with the OSCE grades. Admission interview ranking of the students to the college showed no correlation with the OSCE grades ($r = 0.03$, $p < 0.05$).

Table 4: 2-tailed Pearson correlation statistics and regression analysis of OSCE grades versus prior undergraduate courses' grades to assess predictive validity.

	PHAR441 Total	PHAR441 SMSA	PHAR440 total	PHAR440 SMSA	Medicinal Chemistry	PH590 ICBL	PH491 ICBL
OSCE grades (R)	.467	.613	.723	.782	.335	.653	.429
Sig. (2-tailed)	**.033**	**.003**	**.000**	**.000**	.138	**.001**	.052
R^2 (%)	0.218 (21.8)	0.376 (37.6)	0.523 (52.3)	0.612 (61.2)	Not feasible	0.426 (42.6)	Not feasible
N	21	21	21	21	21	21	21

2.3.3.2 Concurrent validity

Results of our study shows that the OSCE grades have no correlation with the Prometric exam ($r = 0.09$, $p > 0.05$), and MCQ component for the final cumulative assessment ($r = 0.03$, $p > 0.05$); however, sub-analysis of the Prometric exam grades shows that there is a significant strong positive correlation with pharmacy practice and clinical pharmacy component of the exam (r = 0.62, p < 0.05). OSCE grades showed moderate correlation with the cumulative GPA scores of graduating pharmacy students (r = 0.46, p > 0.05), as shown in table 5.

Table 5: 2-tailed Pearson correlation statistics of OSCE grades versus different assessments' grades to assess concurrent validity.

	Cumulative GPA	Prometric Grades	Prometric - Pharmacy Practice and Clinical Pharmacy	PHAR590 MCQ
Pearson Correlation	.456	.092	.618	.032
Sig. (2-tailed)	.101	.755	**.019**	.892
N	14	14	14	21

2.3.4 Internal validity

Correlation between analytical checklist and global scoring per assessor varied significantly. For instance, for assessor 18, there was a very weak positive correlation between his/her global and analytical evaluation (r = 0.1, p > 0.05); on the other hand, assessor 8 showed a significant strong positive correlation between his/her global and analytical scoring (r = 0.76, p < 0.01). Average correlation of all assessors' scoring that participated in the OSCE was (r = 0.52) as shown in table 6.

Table 6: Internal validity assessed through Spearman correlations between the analytical checklist scores and global scores of each assessor were conducted to assess the risk of bias; average of all correlations was calculated. (*: significant at alpha level 0.05; **: significant at alpha level of 0.01).

Assessor	1	2	3	4	5	6	7	8	9	10	11
R	.649**	.323	.493*	.178	.563**	.563**	.682**	.757**	.624**	.721**	.469*
Assessor	12	13	14	15	16	17	18	19	21	22	Avr.
R	.440*	.381	.565**	.371	.631**	.699**	.104	.415*	.532**	.702**	.517

2.3.5 Content validity

Eighteen students were administered the questionnaire and 18 students replied (response rate = 100%). All respondents were female graduating pharmacy students who participated in the OSCE. Degree of satisfaction and resemblance to real practice of individual stations varied significantly from 30.6% (station 6) to 100% (stations 2 & 4), (Table 7). Average satisfaction of all stations reached 73%. Individual comments for all stations are summarized and presented below:

1) Station 1: 3 students disagreed with route of the drug administration, where they claimed that it is mainly topical in the Qatari practice.
2) Station 2: 1 student commented on the low quality of the SP performance

3) Station 3: 4 comments were received. One student was confused about the aim of the station, other student thought that it did not really reflect real practice, and two students complained about the physician's performance.
4) Station 4: No comments were received.
5) Station 5: 4 individual comments varied from the need for knee examination, patient role clarity, rarity of using amlodipine with edema, or disagreement with the fact that patient comes to the pharmacist in the first place.
6) Station 6: 6 students commented mainly on job description and the common sense in practice regarding technician roles and understanding from regulations
7) Station 7: 5 comments varied between insufficient amount of time, comprehensiveness of the case, or patients coming with the insulin or asking about its administration.
8) Station 8: 3 comments were received regarding ethical concern, case clearance, patient cooperation, or fitting of the role for the physician more the pharmacist.
9) Station 9: one student doubted the correctness of the regimen.
10) Station 10: one student felt that the station is not common in the practice.

The results of the SWOC analysis will be discussed in Chapter 3.

Table 7: Scores based on (Yes/No) answers of 18 students participated in the mini questionnaire. Average of stations' satisfaction is calculated.

Station	Students Summed Score per Station	%
Station 1 (Valacyclovir)	14	77.78
Station 2 (Eye Drops)	18	100
Station 3 (Chemotherapy Dose)	12.5	69.44
Station 4 (Diarrhea)	18	100
Station 5 (Swollen Ankles)	12.5	69.44
Station 6 (Technician)	5.5	30.55
Station 7 (Diabetes teaching)	12	66.67
Station 8 (Lorazepam)	10	55.56
Station 9 (Azithromycin)	13	72.22
Station 10 (Contraception)	16	88.89
Average		73.06

2.3.6 Inter-rater reliability

For the analytical ratings, all stations showed significant excellent ICC between raters (assessors) ($p < 0.01$); the lowest station ICC was 0.8 (0.426 - 0.92), station 8, and the highest was 0.94 (0.86 - 0.972), station 7, which was confirmed by Pearson's correlation ($r = 0.74$, $p < 0.01$), ($r = 0.89$, $p < 0.01$) respectively. For global rating, stations showed great variability. Excellent ICC was determined at alpha level of 0.01 between raters in station 6, (0.897 (0.773 - 0.954)) to poor in station 3 (0.176 (-0.492 - 0.588)). These results were confirmed by the Spearman's correlation as shown in table 8.

Only 2 (20%) stations showed poor ICCs in global rating, which are station 3 and 8. The rest of stations showed either fair to excellent reliability (Table 8).

Table 8: Inter-rater reliability using single-measure two-way random ICC, and Pearson's and Spearman's correlation of raters scoring in individual stations for 2015's cycle. (*: significant at alpha level 0.05; **: significant at alpha level of 0.01).

	Analytical score ICC (CI)	Global score ICC (CI)	Analytical score (R)	Global score (R)
Station 1	0.891** (0.756-0.951)	0.864** (0.701-0.939)	.798**	.760**
Station 2	0.915** (0.803- 0.962)	0.563* (0.009-0.805)	.857**	.372*
Station 3	0.84** (0.644-0.928)	0.176 (-0.492-0.588)	.720**	.121
Station 4	0.877** (0.716-0.946)	0.661* (0.241-0.848)	.811**	.542**
Station 5	0.932** (0.789-0.975)	0.672** (0.346-0.852)	.819**, .901**, .929**	.503**, .411*, .381*
Station 6	0.903** (0.713-0.961)	0.897** (0.773-0.954)	.867**	.749**
Station 7	0.937** (0.86-0.972)	0.332 (-0.437-0.696)	.885**	.154
Station 8	0.8** (0.426-0.92)	0.719** (0.382-0.873)	.740**	.646**
Station 9	0.89** (0.756-0.95)	0.506* (-0.023-0.771)	.810**	.299
Station 10	0.827** (0.611-0.923)	0.742** (0.429-0.884)	.697**	.554**
Average	0.8812 (0.7074 - 0.9488)	0.613 (0.1929 - 0.821)	0.755	0.422

2.3.7 Internal consistency

Using Cronbach's alpha to determine the internal consistency of students' performance in the OSCE, the exam stations showed good internal consistency for global rating (Cronbach alpha = 0.874). It also showed excellent internal consistency in students' whole (analytical plus global) performance in all stations (Cronbach alpha =

0.927). Sub-group analysis showed slightly improved internal consistency of undergraduate pharmacy students, (Cronbach alpha = 0.876) for global scoring, and (Cronbach alpha = 0.932) for total grades. Part-time PharmD students showed unacceptable internal consistency for both measures (Cronbach alpha < 0.5) as shown in table 9.

Table 9: Internal consistency of students' performance in OSCE exam in terms of total and global scores using Cronbach's alpha, with sub-analysis of undergraduate students and PharmD students groups.

	Cronbach's alpha	N
Global Score (All Candidates)	.874	26
Total Score (All Candidates)	.927	26
Global Score (Undergraduate Students)	.876	21
Total Score (Undergraduate Students)	.932	21
Global Score (PharmD Students)	.072	5
Total Score (PharmD Students)	.212	5

2.3.8 Analytical checklist's item revision

When the analytical checklists' items were reviewed against student performance, 11 out of 136 (8%) items were found to be scored less than 10% by students. These 11 items are identified in table 10. Station 9 showed the highest number of items that have been scored less than 10% by students (4 out of 11), with the majority of items that needs to be revised were identified in stations 6, 8, 9, and 10. No trends or patterns between revised items were observed.

Table 10: Analytical checklists' items of the lowest score in the whole OSCE exam

Station name	Items of the lowest score	% of checked scores
Station 3 Breast cancer	Asks about patient's CBC with differential	1.92
Station 6 Pharmacy technician	Suggests to circulate a memo with the regulations in relation to roles and duties of pharmacy technicians to all pharmacy technicians employees	7.69
Station 6 Pharmacy technician	Confirms the technician's understanding of his role	9.62
Station 8 Ethics	Asks why patient has needed early refills	3.85
Station 8 Ethics	Explains patient is refilling prescription too early	9.62
Station 9 T.B	Asks about travel history	3.85
Station 9 T.B	Makes referral to physician for Tuberculosis work-up	7.69
Station 9 T.B	Identifies tuberculosis as a possible diagnosis	9.62
Station 9 T.B.	Explains alarm symptoms of tuberculosis (e.g. cough, low grade fever, night sweats and significant weight loss)	3.85
Station 10 Pills	Asks if patient is using any other contraceptive methods (e.g. barrier method)	9.62
Station 10 Pills	Elicits that patient has not yet restarted her oral contraception	1.92

2.4 Summary

The results obtained for this part of the analysis demonstrate acceptable validity and reliability for the 2015 OSCE, as discussed in Chapter 4. Although not a specific objective of this project, results for inter-rater reliability were improved in the 2015 cycle. The ICC of analytical scoring increased from 0.77 in 2014 to 0.88 in 2015, the same applied to the global rating (0.48 to 0.61) and the total ICC (0.64 to 0.75). Interpretations of the results obtained for each of these analyses will be discussed further in Chapter 4.

Chapter 3: Critical Evaluation of the 2015 OSCE: A Qualitative Analysis

Definitions

SWOC analysis: It is also known as SWOT analysis. It is strategic planning instrument that can be utilized at overall organizational level to review a process and make a decision; an analysis of strengths, weaknesses, opportunities, and challenges or threats (107).

3.1 Introduction

This chapter describes the methods and results pertaining to our second research question. Using qualitative methodology, we conducted a critical evaluation of the OSCE organized in 2015 at the College of Pharmacy, Qatar University.

3.1.2 Research Question

The OSCE has been adopted and conducted in Qatar in the CPH-QU for the second time in 2015. As previously discussed, this iteration incorporated recommendations to strengthen the validity and reliability of the exam. The past chapter demonstrated acceptable psychometrics obtained this iteration and this chapter provides more critical feedback for future refinements to further enhance exam credibility. The research question was ‘how can a performance-based assessment in the gulf setting be further refined to meet international standards of validity and reliability.’ We answered this research question using qualitative methodology, as described below.

3.1.3 Objectives

The purpose of this study was to critically evaluate the OSCE and to generate recommendations and theory regarding future assessment implementations through critical analysis of the OSCE from stakeholders' perspectives". This study also contributed to assessment of exam validity.

3.2 Methodology

3.2.1 Study design

3.2.1.1 Qualitative analysis

A qualitative research design using focus groups was chosen because it is exploratory in nature and can provide greater understanding of meanings, reasoning, and opinions as compared to quantitative data. The nature of qualitative research is based on in depth exploration of the research problem, which is difficult to solve by quantitative research (108). Applying qualitative research can provide hypotheses or recommendations for future quantitative research (109). Based on these reasons, we chose to supplement that quantitative psychometric analysis with a qualitative analysis to better understand results and to generate recommendations for future refinements to better exam credibility.

3.2.2 Sample selection

3.2.2.1 Purposeful sampling

Purposeful sampling was used to select participants for this study. Purposeful sampling is a non-randomized sampling technique where the researcher takes control of choosing the sample to participate in the research based on preferred characteristics or roles (110, 111). This sampling technique was used due to a limited population to choose from and the need to select stakeholders with differing roles throughout the exam process.

3.2.2.2 Sample criteria

A sample was chosen from the population of stakeholders who participated in the second OSCE cycle. We chose different type of participants, including administrators, candidates, assessors, standardized patients (SPs), and exam center staff to gain a comprehensive appreciation of the exam from differing perspectives. Some of the participants were purposefully chosen because they were involved in the whole OSCE process from the very beginning of case writing passing through the training until the day of the assessment itself.

3.2.2.3 Selection process

The participants selected by the researchers were individually invited through emails to participate in the study at preselected dates and times. The introductory emails provided them with the purpose of the study, the venue, and the procedures being used for the focus group study.

3.2.3 Data collection

3.2.3.1 Purpose of choosing a specific design, focus groups method

Focus groups were chosen as the methodology to answer our research question. We wanted to have opinions and thoughts from the perspective of exam participants on the points of strengths and weaknesses of the current OSCE, factors that can improve future OSCEs and/or challenges that can threaten such an assessment. This design allowed for generation of thoughts and opinions from varying stakeholders (SPs, assessors, students, and administrators) in the same place, which fostered constructive interaction.

3.2.3.2 Focus groups

Focus groups can be defined as group interviews. Typically, a focus group is a meeting that hosts a number of interviewees, normally from 6 to 8. One or 2 moderators are usually responsible for directing this meeting in order discuss a certain topic or topics (112). These meetings are repeated with similar samples, in order to ensure all thoughts, ideas, and perspectives are documented. Once no new information is provided, the process can be stopped and this point is called the point of saturation (113). Moderators can have the group interviews structured with very specific questions to ask for all attendees, semi-structured where questions are used only for guidance in the meeting, or unstructured where attendees are responsible for all the discussion during these meeting (114).

3.2.3.3 Focus group process

In our design, we chose an unstructured-focus-group approach, where we emailed the attendees with the topic of discussion in the focus groups so they can come prepared to provide their thoughts and discuss them with others participating in the same focus group.

At the selected dates and times, 2 facilitators were prepared to moderate the focus groups together, one of them had previous experience in moderating focus groups and interviews, and the other one received a prior training with an experienced researcher. Both facilitators had no major role in the development and implementation of the OSCE exam but were known to participants. In each focus group, the attendees were asked to sign a confidentiality agreement at the beginning of the meeting to prevent spreading the content of the meeting. During the meeting, they were also provided with food and beverages. Participants were informed that the focus group would be audio-recorded using a "Samsung Note 3 mobile device" for the whole meeting duration. The moderators introduced themselves and confirmed the purpose of the meeting before starting. The definition of the SWOC analysis in the context of the OSCE was given, which was as follows:

1) "S" stands for strengths, where the participants were required to mention the points of the strengths they found in the second OSCE cycle conducted in 2015 from their own perspective.
2) "W" stands for weaknesses, where they were required to mention the points of the weaknesses they found in the second OSCE cycle conducted in 2015 from their own perspective.
3) "O" stands for opportunities, where the participants were asked to provide ideas and thoughts from their opinions and experience about the factors that can help future OSCEs to succeed and improve compared to the current OSCE conducted.
4) "C" stands for challenges, where they were asked to identify risks that can threaten success of future OSCE exams based on their experience.

After giving them this quick introduction, the attendees were responsible for initiating discussion. At this point, the moderators' roles were limited to the following: one moderator was responsible for confirming the points discussed and writing them under the right category on a whiteboard and the other moderator was responsible for taking summary notes about what was discussed by each participant during the meeting. Both moderators were also responsible to keep the participants focused in their discussion in order to completely analyze the OSCE according to the 4 categories. Video tapping of the focus groups was not needed because nonverbal communication and facial expressions would not add to study objectives. By the end of each meeting, photos were taken for the themes created on the whiteboard then transferred to a password-protected University laptop along with an audio file of the recorded-focus-group. Focus groups lasted on average for 1 hour.

3.2.4 Data analysis

3.2.4.1 Transcription

Transcription was the first step in initiation data analysis (115). In order to provide reproducible analysis, we converted the audio data to a written format. The researcher used different instruments to facilitate the transcription process. The main tool was a free transcription web service (116), where the researcher uploaded the audio file and controlled the speed of audio playing. Using the help of the summary notes created during the focus groups, the researcher transcribed the content discussed in the focus groups verbatim in the site’s blank sheets, and then copied the text to a word document file in order to use for further analysis.

3.2.4.2 Content analysis

The qualitative analysis was done manually by two researchers. The researchers conducted a content analysis for the transcriptions created. Content analysis is an instrument used in research in order to detect the existence of specific concepts or words within written data (117, 118) such as transcripts of interviews or focus groups or even whole books. The data detected are further analyzed and categorized. In this research, one researcher began the process of content analysis by using data generated on the whiteboard during the focus groups. A coding framework was developed from the photographs based on the four major categories (Strengths, Weaknesses, Opportunities, and Challenges). Individual codes were the words written under each of the major categories. Then, the researchers applied the coding framework to the focus group transcripts. During this process, each phrase or thought was separated and coded according to one of the codes in the framework. If a code did not exist that captured the phrase or thought, a new code was developed under the relevant category. This process was repeated and transcripts reanalyzed until a final stable framework emerged. The final themes and codes were given to another researcher along with the coded transcripts in order to check the accuracy of coding. Any disagreement in coding was resolved through discussion.

3.2.5 Validity and reliability

To ensure the validity of our SWOC analysis, we incorporated some validation measures. First, we did not invite any of the chief examiners who were responsible for coordinating and organizing the OSCE exam as focus group participants. Their presence could have biased the analysis to be in favor of the exam process either directly through their comments and discussion or indirectly through their presence, which may censor ideas from the other attendees. Second, the moderators were not allowed to give opinions or participate in discussions in order to preserve the internal validity. Third and last, the final content analysis was checked for accuracy by an independent investigator.

3.2.6 Researcher bias

The main researcher of this study was not involved in the OSCE organization and coordination. He did not participate as a standardized patient, assessor, or administrator. Therefore, his analysis of the OSCE was based on the data he received without any subjective opinion.

3.2.7 Ethical approval

The project was approved by Qatar University, QU-IRB 373-E/14.

3.3 Results

3.3.1 The process, focus groups and transcription

We conducted 2 focus groups with total attendees that were selected by purposeful sampling reached 14 volunteers, a response rate of 100%. The attendees included 4 faculty assessors, 1 external assessor, 3 standardized patients, 3 exam center staff, and 3 undergraduate pharmacy students. A total of 103 minutes were recorded in both groups that resulted in 37 pages of transcription. After conduction of two focus groups, it was deemed saturation was reached and no further focus groups were planned or conducted.

3.3.2 Main themes

The focus group discussion resulted in 20 main themes that are distributed in the 4 main categories. Strengths included training, assessment, familiarity, standardization, and satisfaction. The weaknesses included discomfort, assessment, exam organization, and training. The opportunities consisted of future licensure, regulator buy in, improvement, SP pool. The challenges category included novelty, failure policies, specialized pharmacists, preparation of practicing pharmacists, collaboration, cultural differences, OSCE overall scoring. Further subthemes and content analysis are described in the next sections.

3.3.3 Coding

Coding resulted in the themes and subthemes that are presented in table 11 below:

Table 11: Themes and codes generated from the content analysis of the focus groups.

Strengths:

1. Training
   1.1. Assessors
   1.2. SPs
   1.3. Students
      1.3.1. Mock OSCE
      1.3.2. Professional Skills Courses
2. Assessment
   2.1. Mutual grading system (analytical and global scoring)
   2.2. Skills diversity
   2.3. Multiple assessors
   2.4. Practice resemblance

2.5. Time

2.6. Collaboration

2.6.1. Case building

2.6.2. Assessment

3. Familiarity

3.1. Assessors

3.2. Resources

4. Standardization

4.1. SP consistency

5. Satisfaction:

5.1. Students

Weaknesses:

1. Discomfort

1.1. Assessors

1.1.1. Refreshments

1.1.2. Rest

1.1.3. Exam duration

1.2. Students

1.2.1. Refreshments

1.2.2. Assessor unprofessionalism

2. Assessment

2.1. Standardization

2.1.1. Door instructions

2.1.2. Resources feasibility

2.1.3. Standardized patients

2.1.4. Case validation

2.1.5. Practice resemblance

2.2. Grading

2.2.1. Fairness

2.2.2. Subjectivity

2.2.3. Checkmark system

2.3. Time

2.3.1. Students

2.3.1.1. Sticker experience

2.3.1.2. Student readiness

2.3.2. SPs

2.3.2.1. Case familiarity

2.3.3. Assessors

2.3.3.1. Case familiarity

2.3.3.2. Reflecting in global assessment

3. Exam Organization

3.1. Insufficiency

3.1.1. Assessors

3.1.2. SPs

3.1.3. Space

3.2. Lack of coordination

3.3. Interaction

3.3.1. Students

3.3.2. SPs

3.3.3. Other personnel

4. Training

4.1. Students

4.2. Assessors

4.2.1. Cultural communication assessment

Opportunities:

1. Future licensure
2. SCH buy in
3. Improvement
    - 3.1. Grading system
        - 3.1.1. SP involvement
    - 3.2. Bell system
    - 3.3. Sticker system
    - 3.4. Training
        - 3.4.1. SP
            - 3.4.1.1. Customized script
            - 3.4.1.2. Instructions
        - 3.4.2. Students
        - 3.4.3. Assessors
            - 3.4.3.1. Cultural communication assessment
            - 3.4.3.2. SP roles
    - 3.5. Exam resources
    - 3.6. Curriculum
    - 3.7. Recruitment
    - 3.8. Exam timing
    - 3.9. Door instructions
4. SP Pool

Challenges:

1. Novelty of OSCE idea in Gulf
2. Dealing with fails
3. Specialized pharmacists
4. Preparation of practicing pharmacists

5. Collaboration
6. Cultural difference
7. OSCE overall scoring

3.3.4 Content analysis

3.3.4.1 Strengths

Differing stakeholders agreed on the strength of the training given to them. Assessors believed that they received a strong training from experts in the field, which made them better than many schools conducting the OSCE exam. Standardized patients believed they received adequate training. Add to that, students felt that the OSCE was similar to the structured-multi-skill-assessment (SMSA) that they take every year in the undergraduate program. Other students mentioned that the Mock OSCE they took was good preparation for the actual OSCE.

> *"I think one of the strengths of it is that we've actually got through the program. We've actually had formal training from somebody who has had experience in this area, so I feel like got the process better than a lot of schools." (Internal Assessor 1)*

> *"It's like the SMSA, so it's not a new thing for us." (Student 1)*

> *"I think of strength… the fact that we had a mock OSCE; we understood the actual set up." (Student 2)*

The OSCE as an assessment showed many points of strengths that was pointed out by the attendees. For instance, students felt that the mutual grading system, where grading communication skills was part of it, was beneficial for them. They liked that the exam was assessing diversity of skills. Both students and assessors liked the fact that such an assessment as OSCE involves multiple assessors in students' evaluation; while students believed it gave balance, assessors believed that it was a good chance to have the students evaluated from those of pharmacy practice background and pharmaceutical sciences background. Standardized patients noted that the pharmacy students were given enough time to read and understand their role before entering to their stations. Some assessors thought that having a collaboration from different institutions including Sidra, Hamad Medical Corporation (HMC), and Qatar University in case writing and assessing the students was a strength of this exam. OSCE as an assessment showed many points strength related to its structure as a performance-based assessment.

> *"I think, the strength, like also, that there was 2 grading systems."   (Student 2)*
>
> *"There is a lot of things that has been assessed like counselling, educating with a lot of things." (Student 2)*
>
> *"When I think of strength… There was more than one assessor. Because not everyone marks the same, so I feel like balance." (Student 2)*
>
> *"So it's good to mix, someone from practice and someone from pharmaceutical science." (Internal assessor 2)*

> *"I think, one more strength is having several collaboration between several institution like Sidra, people from Sidra, people from HMC, people from QU, so having all this collaboration. That's a strength." (Internal assessor 3)*

It also appeared from pharmacy students' responses that familiarity was important for their performance in the OSCE. They mentioned that familiarity with assessors from the college of pharmacy provided relaxation. They also pointed the importance of being familiar with the book resources that they used in the exam as a reference.

> *"I think it was nice seeing familiar faces when you enter the room." (Student 2)*

> *"We know what resources we are going to use because Dr. "X" posted the name of the books, so we understood that, so it wasn't a total surprise." (Student 1)*

Although it is good to have backup, it was interesting that standardized patients pointed out the importance of being the only one responsible for their stations, where consistency flows with assessing more students.

> *"It was better. It was kind of consistent. I knew what to expect, what to say, and I think also that the assessors said it was good." (Standardized patient 1)*

Pharmacy students showed satisfaction with the exam. They believed it was not stressful.

> *"I think it wasn't stressful." (Student 1)*

3.3.4.2 Weaknesses

The second category was "OSCE weaknesses". Attendees pointing a wide variety of weaknesses in the OSCE process. First, both students and assessors mentioned that they were not feeling comfortable. Some reasons were common such as the need for refreshments because the exam duration is too long for both assessors and students. In addition, assessors complained from the length of the exam; they needed rest or a backup system like standardized patients, in which assessors could substitute between each other while grading students. On the other side, some students believed that some assessors showed unprofessional behavior, knocking on the desk for instance, which distracted them during performing their tasks.

> *"In our room… had no A.C. Students were bothered, assessors were bothered. That affected the performance of some students." (Internal assessor 3)*

> *"…but like refreshments. When we are in the stations, we were talking for about two and a half hours. There is no water." (Student 1)*

> *"I guess even for like SPs, they take turns going in and out. Assessors are there for all eight hours. It's very tiring for us as well. I think we should have like backup system. We should have a break too." (Internal assessor 1)*

> *"… and I think every time we are adding new number of stations. Maybe by next year we will have, I don't know, a whole day. So it will be exhausting for people who are doing this." (Administrator 1)*

> *"Some students were asking for water. There were problems finding water, and I took some water from the dean's office." (Administrator 1)*

> *"… and I think the assessor is bored or something. I was talking. He was sitting on the side, he was bored. He was like (knocking on the table)" "…or someone smiles, we know that we said something stupid." (Student 1)*

For the OSCE assessment itself, it was stated to needed standardization in more than one aspect. For some stations, roles of students should have been stated more clearly. Students mentioned that the resources they used in the station should be electronic to reflect the practice since it is easier and faster. Third, many participants noted that standardized patients should be better standardized; some have extra or less knowledge of their roles depending on their background. Some of them are not consistent in providing the information to students. Fourth, participants also pointed out some weaknesses with case validation, as some cases did not seem to run as planned. In addition to that, it was felt some cases did not necessarily reflect real practice.

> *"And also for the weaknesses, the thing on the door should be more clear ... is our role, like for the manager." (Student 1)*

> *"It wasn't as fast and doesn't really practice when we have our phones or IPADs... And the books mislead us sometimes." (Student 1)*

> *"One of the things that I realized in the station that I was assessing in, we had 2 SPs and, you know, people are different. We don't have much time to memorize and practice the case. So, one of them was able to grasp everything, memorize all the numbers and the blood pressure and everything and the other one after a while started making mistakes… Another thing, that one of them was more, you know, he really wanted the students to do well." (Internal Assessor 2)*

> *"The guy I was with, he is never like even been in the hospital. He know nothing about. He had to play pharmacy technician. So they asked him a lot of clinical questions and role questions… He has never seen what a pharmacy technician*

does. So I think the health care provider roles should be actually played by health care people." (Standardized patient 2)

"For the standardized patients, some of them might, I know it's hard to be consistent, but some of them will give like different information or extra information or something." (Student 1)

"Sometimes they would offer too much more than what was needed ... maybe because they were either pharmacists or health care providers. They weren't really playing the role of the patient." (Internal assessor 1)

"It would be very difficult for a student to have guess night sweats and all the other things if I said my only symptoms was fever." (Standardized patient 3)

"I mean either refer the patients or give the medication. This is my job here, but further diagnosis in community pharmacy!!! It's not common here." (Student 3)

"Like I've been coughing blood, we will think about T.B., but twenty one students, no one realized it was a T.B., because this information did not come up." (Student 1)

"...when we develop the case and then we have to give to another group and then acted it out. I don't know for always 100% into that, we are not, and this why we have a lot of mistakes, and that isn't, we don't realize that until after the day of the exam." (Internal assessor 1)

"I think that maybe the validation should be with an SP." "The SPs ... we invite for the validation. It'd be different. Different SPs for the actual testing, you know,

> *so for the validation part, we use a different SP. we invite separate ones." (Internal assessor 4)*

> *"We are in Qatar, we went on rotation, some of the recommendations that they usually make is not similar to what we had to do here." (Student 2)*

The analysis of grading resulted in 3 subthemes: fairness, subjectivity, and checkmark system. From the comments below, we can see that grading of both content and global skills affected the focus of assessors on evaluating the global skills. Also, students perceived some assessors to be poor communicators themselves, which may have an impact on global assessments. The checkmark system, which is act of assessors checking off points when students achieve them, also appeared to distract students and affected their confidence either positively or negatively.

> *"As an assessor on one of the cases, I was focused a lot on the checklist, analytical checklist, just trying to make sure that, you know, I didn't miss something, you know that student said that I have to tick it. So so much focus there and then at the end you go to the global assessment, which is communication and you kind of feel, you know, I wasn't really focused on that, that much, and I don't have much time. The runner is coming in to take the sheet, so I think we had to do this so quickly and we didn't have much time to reflect." (Internal assessor 2)*

> *"I think that there should be a third person in the room monitoring the person, what they are supposed to say, and checking off if they miss something because it's unfair to the students." (Student 2)*

> *"If I am evaluating, for example, if the student is sitting and doing this (the scene was crossing legs), to me, that's not good and it's negative communication. In other cultures, they are just being comfortable." (Internal assessor 2)*

> *"We don't want to see how many check marks we got." "Some of them, when I tried to know what is on the paper, there is no check mark there; what did I do wrong??" (Student 2)*

> *"I am sure it's stressful when you are looking at us and we are like marking you guys." (Internal assessor 1)*

Under the theme of Assessment, "time" emerged as a major subtheme. It was mentioned by the students, standardized patients, and assessors. Students mentioned their struggle in the station itself, where they had to give coded stickers to the assessors to place on their grading sheets before reading instructions and interacting with standardized patients. Standardized patients believed that they need more time to be familiar with their cases during the dry run. Assessors also reflected on time but in the context of requiring more time to better evaluate global performance.

> *"The sticker issue, that was really bad." (Student 1) "It's time consuming because you walk, and you are getting the… by the time you see the patient walking." (Student 2)*

> *"I think that we just need a little bit more time to settle down because I think the whole process, by the time we settle down, like I find the patient. I am sitting and they are coming in so you don't get to see what's around you. You don't get the chance to comprehend." (Student 2)*

> *"…because especially the management part, there was lots of information. I just started talk with the guy one, and then, he was talking, I was trying to read what's going on, on the table, 3 papers there." (Student 3)*

> *"But I think we need some more time for practice, to get to know the case." (Standardized patient 1)*

> *"Some more time should be spent in orienting the SPs, between the assessors and the SPs so that they know the case very well to know how to answer" (Internal assessor 3)*

> *"That's the orientation between the SPs and the assessors to dedicate more time so that to assess all communication" (Internal assessor 3)*

Other organizational factors such as recruitment, space, and bell alerts were deemed to be weaknesses.

> *"I think recruitment this year was little bit (issue) …SPs and, I don't know, assessors maybe." (Internal assessor 1)*

> *"I felt that one weakness was the location and the space where the exam took place because I felt that there should be like a specialized area for such stations." (Internal assessor 3)*

> *"I think there was a lack of coordination between the 2 sides, between the different sections, like when they ring the bell, then they will knock a bit later." (Student 2)*

Below the theme of training, there were 2 main subthemes that would be considered as weaknesses in the exam. Students believed that they lacked enough training on managerial roles in pharmacies (a competency assessed this cycle), while the rest of participants were focused on communication assessment from different cultures, where they felt cultural communication assessment would be a weakness that could be addressed in the future in order to unify the way of assessing students.

> *"The manager station. I don't think we have enough, that we don't have enough experience for training." (Student 1)*
>
> *"But I think the validity is only in question from a cultural content, right?!" "... I mean there are a lot of cultural issues to take." (Administrator 2)*

3.3.4.3 Opportunities

Through the focus group, many opportunities emerged such as the possibility of using the OSCE for future licensure, the support of the Supreme Council of Health, and the shared pool of standardized patients between different colleges or programs. Also, it was mentioned that some weaknesses could be turned into opportunities if improved. Some of the participants added solutions such as using standardized patients as assessors of communication skills or using within-station bells to solve the problem of the bell system. Others proposed creating customized scripts for standardized patients to decrease their confusion, providing greater training on management competencies, conducting cultural communication training, training individuals to do both roles of assessor and standardized patients to solve recruitment problems, and providing multiple resources in every station such as iPads and clinical books.

"This could be possibly licensing exam in the future." (Internal assessor 1)

"I think one of the strength is that we've buy in from the supreme council of health." (Internal assessor 1)

"I think they will probably do a better job than the assessors. Yea, because, you know, you're communicating to me, it's about how I felt now, not how someone was thinking how I felt through this communication" (Internal assessor 2)

"I think maybe we can have a louder voice" (Administrator 3) "Something within the room itself." (Student 1)

"So should be, like a script basically, a script for the SP that is separate from the development of the case. And, I think it will create less confusion for the SP and the SP will know exactly what is allowed and what not." (Internal assessor 4)

"... and also a supervisor because many pharmacists are supervising other people's work, right?! So we need to kind of provide that education I guess." (Internal assessor 4)

"We will have like a communication course, like two hours, three hours, you know, communications course. We are all on the same page, this is good communication, and this is bad communication. This is ideal, this is not." (Standardized patient 3)

"... Like have cross training, like have our training as standardized patient and also as an assessor. So when they come they can actually do both, versus like, I was only trained to do this because this what I could do. I think you were only

assessors. You know how to do. If we had more flexibility between the two positions, then it can be an opportunity I guess." (Standardized patient 2)

"We could use two different resources in each of the stations." (Internal assessor 1)

"I think that's an important exam, and for us, maybe it's looking at the results and *saying okay, so if some of the students did those mistakes, so what mistakes did we do in the curriculum? What can we go back in the curriculum and make sure we can improve?!" (Internal assessor 2)*

"SPEP rotation, like going around so change these rotations to last semester and then have OSCE at the end, then it makes sense" "then we have the OSCE will be more valid. I mean you can have 50% on OSCE, makes sense, because you have whole semester practice on it, because you are in SPEP rotations." (Student 3)

"Maybe the opportunities, it will be good to do like more detailed instructions, like what we do for SMSA now. We do what the (role) of the student want be, the time, and the name of the patient. And it's like very general instruction." (Internal assessor 1)

"I think all of the colleges share a pool of SPs" (Administrator 2)

3.3.4.4 Challenges

Participants described numerous challenges. The novelty of the idea of OSCE in gulf countries was identified. Assessors were unaware of policies for dealing with failed candidates. If the OSCE was adopted for practicing pharmacists or entry-to-practice, they believed the OSCE would fail some specialized pharmacists even if they are excellent in their departments. It would also be challenging for practicing pharmacists, where information are not fresh in their minds like graduating students. Some participants believed that collaboration in case writing could result in disagreements based on the training background received and could alter case validity. Culture differences would remain a challenge especially when communication evaluation is considered a fundamental part of the OSCE assessment.

> *"The fact it is new, like OSCE is new to the country" (Student 2)*

> *"We keep saying high stakes, but if they fail the OSCE, what happens?!! I mean like, what's the purpose of the OSCE?" (Standardized patient 3)*

> *"Because it's like a high stake exam, if students fail, what would we do? Do we run another OSCE for them?" (Internal assessor 1)*

> *"…because if someone working in the NCCCR in heart hospital, it's hard sometimes to go and tell them to do something this general, like I was working there for 10 years and then you are asking me, okay, it's the basic information but sometimes … "Even if I fail it, this does not mean I am a bad pharmacist in the heart hospital or in the cancer hospital. It depends, I think you should design stations or something that related to their practice. I am not sure if it can, even if for licensing" (Student 1)*

> *"So then, if we are referring this. If this is something that we will roll out to all the practicing pharmacists in general, honestly, I don't think they will pass our*

OSCE. They need a lot of preparation prior to this be even used for a licensing exam. And who is gonna do all this prep work?!!" (Internal assessor 1)

"If I were to do this, I will have to do a lot of studying. I graduated so many years ago" "it will be very challenging for those who are currently practicing" (Internal assessor 1)

"When we were in the case writing session… We had some disagreements about what the students should know… We want to the student to be aware, to ask about platelets… Nobody in my station thought there is a need to know… I fought so strongly about it… I don't know if that's because of where I am trained versus what somebody else's trained or if I am not as familiar with others about the curriculum here and what's taught exactly" (Standardized patient 3)

"I could watch that communication and say they didn't show respect. For you, it was fine; from my culture point of view, it was good, they could impress them… whatever… There is a culture issue here" (Internal Assessor 2)

3.4 Summary

The results of this chapter provide a comprehensive analysis of exam factors by key stakeholders. Findings will support future cycles by continuing to address strengths and opportunities, while focusing on refinements to improve weaknesses and overcoming challenges. The key points from this analysis are presenting in Chapter 4.

Chapter 4: Discussion, Limitations, Conclusion, and Future Studies

4.1 Discussion:

This thesis provided a comprehensive evaluation of the 2nd iteration of a cumulative OSCE implemented for graduating pharmacy students in Qatar. The OSCE was successfully implemented for the previous 2 years in CPH-QU. This OSCE, which was adopted from the Canadian context, was exclusively analyzed quantitatively and qualitatively in order to determine its appropriateness as a high stakes exam and to identify refinements required to improve validity and reliability within the Gulf context. Our aim was to answer our 2 major research questions:

1) "What is the validity and reliability of a 2nd iterative cycle of a cumulative, summative OSCE for graduating pharmacy students in a GCC context?"
2) "How can the OSCE be further refined to improve validity and reliability within the GCC context?"

To answer the first research question, we analyzed the different components of the psychometric analysis:

The first point of consideration relates to standard setting and student pass rate. For this cycle, the borderline regression method (BRM) was used to determine cut scores for each station. This method was chosen based on previous concerns using the Angoff method in our context and it was also convenient, did not consume a lot of resources, and saved time due to its post-hoc nature. In dentistry, this examinee-centered approach was tested and compared with other methods such as borderline modified group method (BGM) and test-centered Angoff methods I and II (Setting standards using judges without reality check after the exam and with reality check after the exam, respectively (119))

(85). It shared similar advantages as the BGM such as the ease of use, the simplicity of statistical calculations, and affordability. Actually, the BRM showed less statistical error in this study compared to the BGM and it was recommended for institutions using OSCE with small number of students due to a potential lack of a sufficient amount of students deemed borderline (85), which was the case in our study. When the BRM was compared with both Angoff methods, it showed better credibility and reliability (degree of statistical error) according to Kramer et al. (119). Credibility was defined as the sensitivity of the method to the difficulty of the station when calculating the cut scores (119). This was in accordance with other studies that showed enhanced reliability, credibility and improved pass rates with BRM compared to Angoff methods (82). Although our study did not compare standard-setting methods directly, we believe the BRM is a suitable approach in our context based on the considerations discussed (available resources, small number of students, time), as well as concerns resulting from past use of the Angoff method. However, we recommend future studies in Qatar and other centers in the Gulf to directly compare these methods to determine any impact on cut scores and student passing rates.

Based on the standards set for this exam, 4 stations failed to have a pass rate more than 80%. This finding was contrary to our expectations that most (if not all) stations would achieve this pass rate benchmark due to the blueprinting of the exam based on minimal competency expectations. Many factors may explain this discrepancy, some of which are discussed below:

1) Students challenging the exam this cycle did not meet competency expectations

2) The pass rate discrepancy may expose validity concerns, especially relating to reflection of real practice in Qatar

3) Inappropriate setting of ‘minimal competency’ (i.e. borderline) on the global assessment as 3/5, instead of 2/5

4) Assessor bias or inexperience

Based on results obtained from both chapters 2 and 3, we believe the most likely explanation for low pass rates on 4 stations to be due to validity concerns, as discussed throughout the rest of this chapter, yet, overall average pass rate of all stations was deemed acceptable (79.2%).

Predictive validity allows us to make assumptions regarding the OSCE’s validity if courses or assessments meant to target similar knowledge and skills associate with OSCE scores. A second type of predictive validity is how OSCE can predict performance of students in their real practice (120-123), which is beyond the goal of this study. Results of our study suggested that the students’ performance in OSCE could be predicted by undergraduate courses (Professional Skills) that develop skills of interacting with patients or other health care providers and other skills such as answering drug information requests. Formative OSCEs (SMSAs) (12) within these courses were also highly predictive of performance on this OSCE. Courses focusing on critical thinking and paper-based problem solving (Integrated Case-based Learning) also were associated with OSCE scores but to a lesser extent. It is important to note that the high predictability of OSCE performance by one of the integrated case based learning courses (PHAR590) could be attributed to the fact that the OSCE exam accounts for 20% of the subject's grade.

Our findings matched what was previously reported in the Department of General Practice in Medicine in two different universities, where they found moderate to high correlation between previously taken written skills tests by students and their OSCE analytical checklist or global scores (124). This was also in accordance with what was demonstrated by Remmen et al.(95).

One key predictive analysis was that of admission interview ranking. Programs around the world are increasingly attempting to select the most likely to be successful students on admission, in order to increase retention and decrease resource consumption for unsuccessful students. Therefore, it was interesting to discover that the admission interview ranking did not predict success on the OSCE. At CPH-QU, admission to the program is largely based on GPA during pre-pharmacy years. This likely means that performance of such assessments is dependent on the knowledge and skills you acquire during the undergraduate program, as opposed to academic success prior to admission to the College. McLaughlin et al. found similar findings regarding to the relation between admission scores and OSCE performance (125). The authors highlighted the importance of academic institutions considering more reliable techniques to assess non-cognitive and professional skills that would be critical for the success of students as practitioners in the future. This finding warrants examination of the current admission process to determine refinements that may better predict success on high stakes performance-based assessments in the future. However, the missing link at this time is whether or not success on the OSCE can be used as a surrogate marker to predict success in practice.

The results of our concurrent validity analyses have major implications for licensure and regulation of pharmacists in Qatar. It was interesting to see that OSCE performance was not associated with overall Prometric exam results. This finding reaffirms an assumption that knowledge and performance-based assessments do not necessarily measure the same competencies expected of pharmacists in practice and is in line with previous studies (14). While we cannot state at this time that the OSCE is a better assessment method for licensure and regulation, it can be speculated that major refinements are required to the current licensing procedures in order to ensure competency of pharmacists is established prior to practice in Qatar. Based on our sub-analysis, however, we found a strong association between the OSCE scores and the scores of students in the pharmacy practice and clinical pharmacy component of the Prometric exam. Here rises a question, if the country is thinking of adapting OSCE on a national level for licensing of pharmacists, should the OSCE replace the Prometric exam or should both assessments be used together? For a comprehensive assessment of clinical competence, it would be wise to use OSCE with other traditional methods (126). Future studies and collaboration with regulating bodies in Qatar should seek to further answer this question.

Another interesting finding was that the OSCE scores did not correlate with scores on the cumulative knowledge-based (MCQ) component. These questions were developed by faculty groups, were blueprinted to the AFPC competencies expected of our graduates (15), and were largely clinical based. Future iterations of these exams should attempt to assess psychometric properties together, in order to improve examination methods as a whole. Based on these findings, it can be strongly recommended that the OSCE is maintained as a cumulative assessment method for the college as it assesses what these other assessments cannot measure, the higher level of learning process, which is to "show how" to use the clinical knowledge acquired, and integrate knowledge with clinical skills and critical thinking in order to solve real practice problems (127).

Our study found minimal risk of bias from assessors (biased ratings on global assessment based on performance on analytical component). The moderate association between analytical checklist and global ratings by assessors was acceptable and in line with other studies. Lila et al. found relatively low to moderate association between global and analytical scoring (128). The authors suggested that both tools cannot replace each other (128) and we agree based on the results we obtained. This moderate association demonstrated that most assessors were most likely able to differentiate in grading using both assessments without relying on analytical checklist evaluation to predict how they score students in the global rubric. This could perhaps be a result of training techniques that directly addressed this point. Therefore, we recommend future training exercises to include this point, as well as provide greater opportunities for recruited assessors to use both tools prior to exam implementation.

The content validity of the OSCE requires further examination. First, as shown previously, the blueprint likely improved its content validity (129, 130). In addition, we observed that the exam received general satisfaction by students in the majority of stations. Although we did not use satisfaction-specific questionnaires like other studies to measure students' satisfaction (131-133), we were able to determine it through: 1) the mini questionnaire that was measuring resemblance of OSCE with real practice, 2) the

SWOC analysis conducted after the OSCE. Of note, no complaints about difficulty or stressfulness of the exam were reported by students, which differed from studies (34, 134-136). Possible explanations could be because the students did not consider the OSCE as a new experience since they completed several formative OSCEs (SMSAs) during their undergraduate level or it could be due to the fact that OSCE grades accounted for only 20% of a two credit-hour course.

The major finding with respect to content validity is student perceptions regarding the exam's reflection of real practice for competencies outside of patient care. In the 2015 cycle, we implemented 1 station to assess competencies related to pharmacy management (station 5). Based on comments provided, students did not accept this station and did not feel if reflected real practice. This finding exposed a curricular gap in performance-based assessment that should be addressed. Specifically, these results must be relayed to College administration to refine curriculum to account for skills required of pharmacists related to non-traditional competencies such as patient care and communication. Competencies such as management, advocacy, and collaboration should be focused on for future skills-based teaching and assessment. Aligning teaching and learning methods with assessment techniques will further improve student perceptions and content validity of this exam.

More work could be done to better match the OSCE the real practice, as the students integrate into practice during their training before attending the exam. However, an important point is that the OSCE also aims to assess competencies according to Canadian standards for accreditation purposes. Therefore, expectations of student performance may be somewhat discrepant with practice in Qatar.

Inter-rater reliability of the exam was deemed to be high for the analytical checklist and moderate for the global assessment. Few studies were identified in pharmacy as comparison, however we believe these results to be strong. While it may seem ideal to target near perfect inter-reliability for both components, it is not appropriate to expect this for global assessments. Each assessor interprets communication and overall effectiveness of interactions in their own way and some discrepancy in inter-rater reliability reflects this. It is possible that in our setting with great diversity in assessor background, culture, and training, it is worthwhile to maintain two assessors per station to account for any bias resulting from an assessor's own preferences that may or may not match the patient's own preferences. The problem, however, may be with certain assessor pairs as reliability differed greatly between stations. For example, one student completing station 5 received a 3, 4, and 5 on global skills from the three assessors in the station. If each one of these assessors was alone, the student may have failed, moderately passed, or almost received a perfect score on the station depending which assessor was present. Therefore, targeted training on assessment and use of tools may be warranted to better standardize global assessments and to avoid these discrepancies in high stakes exams. Other options could include using different raters solely focused on rating the students using the global scoring or the use of standardized patients to evaluate performance. As assessors stated difficulties in completing both analytical and global scoring during the interaction, these could both be valid alternatives.

Comparing our results to the results of OSCE implemented the year before (70), we can see that there is numerical improvement in the interrater reliability, where analytical rating ICC increased from 0.77 to 0.88 and the global rating increased from 0.48 to 0.61. This improvement could be attributed to chance alone, more training allocated to assessors, refinement of the global assessment tool, and/or the experience and familiarity gained by most of the assessors who participated in the first cycle in 2014. Using design-based research methodology to apply further refinements may result in a further improvement for coming cycles.

Internal consistency of the OSCE for both overall and global performance was deemed to be high. However, this finding was mainly attributed to the performance of the undergraduate students and poor internal consistency was noted for the PharmD students. This variance could be explained by the low sample size for PharmD students (n=5), as studies suggest low numbers may affect Cronbach's alpha (137). Other possible reason could be that the part-time PharmD students are already working and specialized pharmacists; their reactions and performance could vary significantly among stations based on their specialty or comfort with case content. However, global assessments should not greatly vary across the entire exam. Due to the fact all 5 of these students were male and all undergraduate students were female, it was known to assessors which program they belonged to and assessor bias of familiarity with the undergraduate approach cannot be ruled out. Implications of this finding could be very significant if an OSCE is to be used for licensure and registration of pharmacists graduating outside of Qatar University, however for the purposes of this evaluation we deem internal consistency for the exit-from-degree model to be strong.

The checklists' items revision was a simple, yet very useful technique to further test the validity of analytical assessment content. It showed the proportion of items in the exam that scored low (<10%) across all students. It also identified stations that were particularly problematic, such as station 9 where 4 of 14 points were achieved by less than 10% of all students. Interestingly, the BRM resulted in a pass rate of >80% for this station, which may mean that the station was overly complex for this exam or standard setting was flawed for this station. We attribute this result to be due to complexity, as results were consistent across students as a whole. This revision process also allowed us to identify points for analysis regarding validity in terms of practice expectations, as well as to identify potential gaps needing addressing in the undergraduate curriculum. However, no pattern in identified items was detected across all stations and so these results should be compared with results from future iterations to determine curricular revision needs.

In order to answer our second research question, “How can the OSCE be further refined to improve validity and reliability within the GCC context?”, we completed a qualitative analysis of key stakeholder (students, assessors, standardized patients, exam center staff) perceptions. Results of the SWOC analysis are comprehensive. It identified many key issues pertaining to many aspects of OSCE design, implementation, and evaluation. As such, an in depth analysis and interpretation of all data is beyond the scope of this project. We therefore decided to focus on three key points that contribute to our understanding and interpretation of results obtained in the psychometric analysis.

The first point relates to content validity. It was signaled from the student perceptions and item revision data described in the psychometric analysis above that some stations may not have reflected current practice in Qatar and/or focused on competencies not addressed within the undergraduate curriculum. This finding supports our interpretations regarding the first research question and allows us to conclude that content validity was not perceived highly for all stations. Specifically, students spoke to a station requiring them to provide remedial feedback for a pharmacy technician. It was identified that this station was not perceived to reflect practice or the undergraduate curriculum as a whole. It is possible that this was a correct perception, or it is possible that this identified a learning gap in the curriculum and/or experiential training activities and that students should be expected to be competent with these skills upon graduation. As this station was blueprinted to the AFPC competencies of Manager (15), we believe the second rationale to be true. Therefore, we recommend assessment of the curriculum and practice site activities to determine how to address this identified need. Also, we recommend future cycles to include stations blueprinted to competencies aside from "Care Provider" and "Communicator", in order to provide more opportunities for identifying curricular gaps and learning needs.

Students mentioned concerns with another station (station 9) regarding content validity. This station required students to assess a patient presenting with a prescription to a community pharmacy for azithromycin and determine that he needed referral back to his physicians due to risk factors and symptoms specific for tuberculosis. While community pharmacy practice is largely underdeveloped in Qatar, patient assessment is a core competency expected of graduates. However, the problem may have been the setting of the case, as this patient likely would have received a prescription from a hospital or clinic. Therefore, we recommend a focus on setting and problem alignment during case validation procedures at the time of case writing.

Training, or preparing different participants for the OSCE was perceived as strong or adequate by different exam's stakeholders. The key success point was attributed to different factors. If we are to discuss it from students' point of view, we would say that such a stressful exam needed a special preparation of students. In other words, the students needed to witness and live the same exam format that depends on performance many times to familiarize them with it, which was the case in the college. They took the exam several times during their undergraduate years but in a formative format (SMSA) and less weight percent. Add to that, to be prepared for a summative type, a mock station was developed to them weeks before the exam. These could be the reasons why the OSCE training was perceived strong by the students. Discussing it from assessors point of view, there were two important factors. They received training from investigators experienced in implementing and running successful OSCEs before. The other factor was that they have been shown all the possible scenarios that could happen by a student in the station (good performer and weak communicator, weak performer but good communicator, good or bad on both) during their training. However, the interrater reliability between assessors on the global scoring suggests that they need more training and focus on this specific instrument. Although the training was perceived adequate by SPs, there were some complains by assessors or students of SPs underperforming or over helping students, which means that either the SPs did not stick to instructions given to them, did not have enough time to absorb their roles, or they were not qualified. More work need to be done regarding SP training.

Different stakeholders considered the OSCE as a successful assessment. The exam maintained the main key factors for its success, which included the reliance on both global and analytical scoring, the use of more than one assessors for evaluation, the evaluation of different skills, the resemblance of practice in most of the cases, and the collaboration between different institutions in setting up the exam. On the other hand, standardization of the assessment, in terms of SPs, cases, and instructions, and proper timing in the process are needed to maintain a more successful assessment context.

Analyzing the findings of the SWOC analysis, generally speaking, it was interesting to see that main themes such as both training and assessment have strengths and weaknesses, which indicated that good work was done to improve the OSCE compared to the cycle before and it also meant that there is still a room for improvement for future cycles. In addition, it is important to say that most of the weaknesses in the exam was identified as opportunities for success of future OSCEs if they could be improved. Same as for students being familiar with OSCE, we would say that a challenge like the novelty of OSCE idea in gulf would change for pharmacists and some assessors within the coming few years and became no longer a challenge. In order to maintain a successful adapted OSCE in the college and even expand it to a national level, it is critical to maintain the strength points that the OSCE already has, tackle all the discovered weaknesses, and most importantly, people working on the OSCE should be thinking one-step ahead, where they have to create solutions to avoid future challenges that could lead to OSCE failure.

Combining all strengths, weaknesses, opportunities, and challenges, we created a list of main points that would be important to consider in order to improve the level of OSCEs implemented and adapted in academic institutions:

- When adapting OSCE for the first time, it is important to include OSCE experts in the process in order to transfer their successful experience.
- Training cannot be a single session prior to OSCE. It should include different stages through undergraduate years in order to familiarize different personnel (students, assessors, and SPs) with the process.
- To have a successful and reliable exam, using the 2 grading instruments (analytical checklist and global scoring), using multiple assessors per station, measuring different skills and learning outcomes, resembling the cases with the real practice, putting clear instructions for assessors, SPs and students, and using collaboration of practitioners and academic staff in the whole process will enhance the success, validity and reliability of the OSCE.
- OSCE can be a long tiring exam so it is important to plan well for such an exam; in other words, the venue should be large enough with sufficient numbers of rooms (stations), the number of rest stations should be adequate, there should be enough refreshments for all participants, and the rooms should be well equipped and comfortable.
- Creating a specialized center of standardized patients that is shared between different health sciences colleges would improve the quality of the standardized patients and the exam itself and would help solve the problem of not finding adequate number of standardized patients.

These considerations could be further analyzed in future iterations throughout a design-based research continuum (138).

4.2 Limitations

The methodology we used for this evaluation was comprehensive, however limitations of our research exist and should be addressed. First, we lacked analyses measuring detailed satisfaction and perceptions of students, standardized patients, and assessors as have been reported in other studies. Instead, we chose a qualitative focus group approach to generate these data. While we believe the focus group approach was appropriate, it may not represent the perceptions of all those involved in the OSCE and sampling all stakeholders in future iterations is warranted. Secondly, we report combined results from undergraduate students (n=21) and PharmD students (n=5) for most analyses. Aside from internal consistency, we did not attempt to stratify results due to the low sample size of PharmD students. Therefore, we cannot make conclusions regarding any potential differences between these mixed pools of candidates. Finally, this analysis was general in nature and was not designed to evaluate specific improvements implemented for the 2015 OSCE or test specific hypotheses based on results obtained from the original pilot. Therefore, we cannot be certain which improvements contributed to positive psychometric results but analysis of the data as described above allows us to make assumptions based on results obtained. Despite these limitations, this analysis provides a comprehensive evaluation of an exist-from-degree OSCE implemented in the Gulf region and gives valuable information regarding validity, reliability, and refinements required for future cycles.

4.3 Future studies

Based on interpretation of our results discussed above, we recommend three key studies for future analysis. First, both the psychometric and qualitative analyses provided signals that stations did not necessarily reflect practice in Qatar, which may have compromised content validity. Therefore, we recommend a more comprehensive analysis of this component in future cycles by measuring perceptions of all stakeholders (students, assessors, and standardized patients), as well as to further analyze checklist points achieved by <10% of students across multiple cycles. Secondly, we recommend studying standardized patient assessments of global performance and comparing to assessor scores, in order to determine if this is a more suitable approach for global assessment in our context. Due to the multicultural and diverse nature of our setting, it is possible the best evaluators of communication skills and overall performance are patients themselves. Lastly, we recommend completing the same analysis in a less homogenous population, as we identified discrepancies between student groups that may or may not affect validity of the exam outside of an exit-from-degree model. Completion of these studies will further develop knowledge and theory pertaining to adaption of performance-based assessment in the Gulf.

4.4 Conclusion

The 2015 OSCE was successfully implemented and evaluated within our context. To answer our first research question, we deem the validity and reliability of this second iteration of the adapted OSCE to be strong. However, future iterations should focus on improving content validity as a whole. With respect to our second research question, we identified many aspects regarding case validity, training, and logistical items that must be enhanced and/or refined to maintain and improve validity and reliability of the OSCE as a high stakes assessment. In conclusion, the 2015 OSCE met expectations as a successful, high stakes, exit-from-degree performance-based assessment.

References

1. Hanna GS, Dettmer P. Assessment for effective teaching: Using context-adaptive planning: Allyn & Bacon; 2004.

2. Fry H, Ketteridge S, Marshall S. A handbook for teaching and learning in higher education: Enhancing academic practice: Routledge; 2008.

3. Bunch MB. Formative, Interim, and Summative Assessments: It Takes All Three 2012 [cited 2015 October, 26]. Available from: http://www.measurementinc.com/sites/default/files/It%20Takes%20Three.pdf.

4. Formative and Summative Assessment: Northern Illinois University, Faculty Development and Instructional Design Center. Available from: https://www.azwestern.edu/learning_services/instruction/assessment/resources/downloads/formative%20and_summative_assessment.pdf.

5. McTighe J, O'Connor K. Seven practices for effective learning. Kaleidoscope: Contemporary and Classic Readings in Education. 2009;174.

6. Interim assessment: The Glossary of Education Reform; 2013. Available from: http://edglossary.org/interim-assessment/.

7. Wiggins G. A true test. Phi Delta Kappan. 1989;70(9):703-13.

8. McMillan JH, Hellsten L, Klinger D. Classroom assessment: Principles and practice for effective standards-based instruction: Pearson/Allyn & Bacon Boston, MA; 2007.

9. Gronlund NE. Assessment of student achievement: ERIC; 1998.

10. Swanson DB, Norman GR, Linn RL. Performance-Based Assessment: Lessons From the Health Professions. Educational Researcher. 1995;24(5):5-11.

11. Wiggins GP. Assessing student performance: Exploring the purpose and limits of testing. San Francisco, CA, US: Jossey-Bass; 1993. xx, 316 p.

12. Miller GE. The assessment of clinical skills/competence/performance. Academic medicine. 1990;65(9):S63-7.

13. Beck DE. Performance-Based Assessment: Using Pre-Established Criteria and Continuous Feedback to Enhance a Student's Ability to Perform Practice Tasks. Journal of pharmacy practice. 2000;13(5):347-64.

14. Wiggins G. The Case for Authentic Assessment. ERIC Digest. 1990.

15. AFPC. Educational outcomes for first professional degree programs in pharmacy (entry-to-practice pharmacy programs) in Canada. 2010 [cited 2016 17 January]. Available from: https://www.afpc.info/sites/default/files/AFPC%20Educational%20Outcomes.pdf.

16. Beck DE, Boh LE, O'Sullivan PS. Evaluating student performance in the experiential setting with confidence. American journal of pharmaceutical education. 1995;59(3):236-47.

17. Watson MC, Skelton JR, Bond CM, Croft P, Wiskin CM, Grimshaw JM, et al. Simulated patients in the community pharmacy setting–Using simulated patients to measure practice in the community pharmacy setting. Pharmacy World and Science. 2004;26(1):32-7.

18. Dick TB, Moorman KL, MacDonald EA, Raines AA, Cox KDM. Defining and implementing a model for pharmacy resident research projects. Pharmacy practice. 2015;13(3):562.

19. Chipman JG, Beilman GJ, Schmitz CC, Seatter SC. Development and pilot testing of an OSCE for difficult conversations in surgical intensive care. Journal of surgical education. 2007;64(2):79-87.

20. Aeder L, Altshuler L, Kachur E, Barrett S, Hilfer A, Koepfer S, et al. The" Culture OSCE"-Introducing a formative assessment into a postgraduate program. Education for health. 2007;20(1):11.

21. Frohna JG, Gruppen LD, Fliegel JE, Mangrulkar RS. Development of an evaluation of medical student competence in evidence-based medicine using a computer-based OSCE station. Teaching and learning in medicine. 2006;18(3):267-72.

22. Fliegel J, Frohna J, Mangrulkar R. A Computer-based OSCE Station to Measure Competence in Evidence-based Medicine Skills in Medical Students. Academic Medicine. 2002;77(11):1157-8.

23. Monaghan MS, Jones RM, Haddad AM, Ineck J. Designing an assessment for an abilities-based curriculum. American journal of pharmaceutical education. 2005;69(1-5):118.

24. Vargas AL, Boulet JR, Errichetti A, van Zanten M, Lopez MJ, Reta AM. Developing performance-based medical school assessment programs in resource-limited environments. Medical teacher. 2007;29(2-3):192-8.

25. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Annals of internal medicine. 2003;138(6):476-81.

26. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Annals of internal medicine. 1995;123(10):795-9.

27. Beck DE, O'Sullivan PS, Boh LE. Increasing the accuracy of observer ratings by enhancing cognitive processing skills. American journal of pharmaceutical education. 1995;59:228-.

28. McDonough RP, Bennett MS. Improving communication skills of pharmacy students through effective precepting. American journal of pharmaceutical education. 2006;70(3).

29. Tugwell P, Dok C. Medical record review. Assessing clinical competence New York: Springer. 1985;142.

30. Challis M. AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education. Medical teacher. 1999;21(4):370-86.

31. Boud D, Keogh R, Walker D. Reflection: Turning experience into learning: Routledge; 2013.

32. Fung MFK, Walker M, Fung KFK, Temple L, Lajoie F, Bellemare G, et al. An Internet-based learning portfolio in resident education: the KOALA™ multicentre programme. Medical education. 2000;34(6):474-9.

33. Austin Z, O’Byrne C, Pugsley J, Munoz LQ. Development and validation processes for an objective structured clinical examination (OSCE) for entry-to-practice certification in pharmacy: the Canadian experience. American journal of pharmaceutical education. 2003;67(3):76.

34. Awaisu A, Abd Rahman NS, Nik Mohamed MH, Bux Rahman Bux SH, Mohamed Nazar NI. Malaysian pharmacy students' assessment of an objective structured clinical examination (OSCE). American journal of pharmaceutical education. 2010;74(2):34.

35. Salinitri FD, O'Connell MB, Garwood CL, Lehr VT, Abdallah K. An objective structured clinical examination to assess problem-based learning. American journal of pharmaceutical education. 2012;76(3):44.

36. Ishikawa H, Hashimoto H, Kinoshita M, Fujimori S, Shimizu T, Yano E. Evaluating medical students' non-verbal communication during the objective structured clinical examination. Medical education. 2006;40(12):1180-7.

37. Harden R, Stevenson M, Downie WW, Wilson G. Assessment of clinical competence using objective structured examination. Bmj. 1975;1(5955):447-51.

38. Tokunaga J, Takamura N, Ogata K, Setoguchi N, Utsumi M, Kourogi Y, et al. An advanced objective structured clinical examination using patient simulators to evaluate pharmacy students' skills in physical assessment. American journal of pharmaceutical education. 2014;78(10):184.

39. Pugh D, Hamstra SJ, Wood TJ, Humphrey-Murto S, Touchie C, Yudkowsky R, et al. A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. Advances in Health Sciences Education. 2015;20(1):85-100.

40. Sakurai H, Kanada Y, Sugiura Y, Motoya I, Wada Y, Yamada M, et al. Reliability of the OSCE for Physical and Occupational Therapists. Journal of physical therapy science. 2014;26(8):1147.

41. Lele SM. A mini-OSCE for formative assessment of diagnostic and radiographic skills at a dental college in India. Journal of Dental Education. 2011;75(12):1583-9.

42. Meskell P, Burke E, Kropmans TJ, Byrne E, Setyonugroho W, Kennedy KM. Back to the future: An online OSCE Management Information System for nursing OSCEs. Nurse Education Today. 2015;35(11):1091-6.

43. Näpänkangas R, Karaharju-Suvanto T, Pyörälä E, Harila V, Ollila P, Lähdesmäki R, et al. Can the results of the OSCE predict the results of clinical assessment in dental education? European Journal of Dental Education. 2014.

44. Brennen W, Carter A, Rudd Prof C, Ross Mr N, Claxton Mrs L. Clinical placement before or after simulated learning environments? A naturalistic study of clinical skills acquisition amongst early-stage paramedicine students. eCULTURE. 2015;7(1):3.

45. Artemiou E, Adams C, Hecker K, Vallevand A, Violato C, Coe J. Standardised clients as assessors in a veterinary communication OSCE: a reliability and validity study. The Veterinary Record. 2014;175(20):509-.

46. Kheir N, Awaisu A, Ndoye A, Wilby KJ. Structured Multi-Skill Assessment (SMSA) in pharmacy: A contextual adaptation for authentic assessment for colleges of pharmacy and beyond. Avicenna. 2015.

47. Han JJ, Park H, Eo E, Yoo K, Lee D, Jung WS. An OSCE for summative assessment after clinical clerkship: experience in Ewha Medical School. Korean Journal of Medical Education. 2004;16(1):33-40.

48. Brailovsky CA, ‘Maison PG, Lescop J. Construct validity of the québec licensing examination SP-based OSCE. Teaching and Learning in Medicine: An International Journal. 1997;9(1):44-50.

49. Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. Academic Medicine. 1996;71(1):S19-21.

50. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Academic Medicine. 1993;68(6):443-51.

51. Dupras DM, Li J. Use of an objective structured clinical examination to determine clinical competence. Academic Medicine. 1995;70(11):1029-34.

52. Pierre RB, Wierenga A, Barton M, Branday JM, Christie CD. Student evaluation of an OSCE in paediatrics at the University of the West Indies, Jamaica. BMC medical education. 2004;4(1):22.

53. Eldarir SA, El Sebaae HA, El Feky HA, Hussein H, El Fadil N, El Shaeer IH. An introduction of OSCE versus traditional method in nursing education: Faculty capacity building and students’ perspectives. Journal of American Science. 2010;6(12):1002-14.

54. Austin Z, Gregory P, Tabak D. Simulated patients vs. standardized patients in objective structured clinical examinations. American Journal of Pharmaceutical Education. 2006;70(5).

55. Howley LD. Performance Assessment in Medical Education: Where We’ve Been and Where We’re Going. Evaluation & the health professions. 2004;27(3):285-303.

56. Cox M, Irby DM, Epstein RM. Assessment in medical education. New England Journal of Medicine. 2007;356(4):387-96.

57. Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. Family Medicine. 2008;40(8):574-8.

58. Zahid MA, Al-Zayed A, Ohaeri J, Varghese R. Introducing the objective structured clinical examination (OSCE) in the undergraduate psychiatric curriculum: evaluation after one year. Academic Psychiatry. 2011;35(6):365-9.

59. Turan S, Konan A. Self-Regulated Learning Strategies Used in Surgical Clerkship and the Relationship with Clinical Achievement. Journal of surgical education. 2012;69(2):218-25.

60. Selim AA, Ramadan FH, El-Gueneidy MM, Gaafer MM. Using Objective Structured Clinical Examination (OSCE) in undergraduate psychiatric nursing education: Is it reliable and valid? Nurse education today. 2012;32(3):283-8.

61. Raheel H, Naeem N. Assessing the Objective Structured Clinical Examination: Saudi family medicine undergraduate medical students' perceptions of the tool. JPMA The Journal of the Pakistan Medical Association. 2013;63(10):1281-4.

62. Karim JA, Marwan YA, Dawas AM, Akhtar S. Self-confidence of medical students in performing clinical skills acquired during their surgical rotation. Assessing clinical skills education in Kuwait. Saudi medical journal. 2012;33(12):1310-6.

63. Idris SA, Hamza AA, Elhaj MA, ElzakiElsiddig K, Hafiz MM, Adam ME. Students' Perception of Surgical Objective Structured Clinical Examination (OSCE) at Final Year MBBS, University of Khartoum, Sudan. Medicine Journal. 2014;1(1):17-20.

64. Eswi A, Samy A, Shaliabe H. OSCE in Maternity and Community Health Nursing: Saudi Nursing Student's Perspective. American Journal of Research Communication. 2013;1(3):143-62.

65. Abdulla MA. Student's Perception Of Objective Structured Clinical Examination (Osce) In Surgery At Basrah College Of Medicine. Bas J Surg. 2012;18:1-6.

66. Abdelaziz A, Hany M, Atwa H, Talaat W, Hosny S. Development, implementation, and evaluation of an integrated multidisciplinary Objective Structured Clinical Examination (OSCE) in primary health care settings within limited resources. Medical teacher. 2015(0):1-8.

67. Zerrin Toklu H. Problem Based Pharmacotherapy Teaching for Pharmacy Students and Pharmacists. Current drug delivery. 2013;10(1):67-70.

68. Johnson B, Pyburn R, Bolan C, Byrne C, Jewesson P, Robertson-Malt S, et al. Qatar Interprofessional Health Council: IPE for Qatar. Avicenna. 2011(2011):2.

69. Al-Azzawi AMJ, Nagavi B, Hachim MY, Mossa OH. The implementation and development of an objective structured clinical examination in the community pharmacy course of a select Gulf-region academic institution (Ras Al Khaimah College of Pharmaceutical Sciences): a pilot study. Innovations in Education and Teaching International. 2013(ahead-of-print):1-13.

70. Wilby KJ, Black E, Austin Z, Mukhalalati B, Aboulsoud S, Khalifa SI. Psychometric Evaluation and Contextual Adaptation of a Final Cumulative. OSCE for Pharmacy Students in Qatar. 2014.

71. Qatar's Statistics World Bank Group [cited 2015 October, 26]. Available from: http://databank.worldbank.org/data/reports.aspx?source=2&country=QAT&series=&period=#.

72. Mahgoub Y, Qawasmeh RA. Cultural and Economic Influences on Multicultural Cities: The Case of Doha, Qatar. Open House International. 2012;37(2).

73. Qatar national vision 2030 [cited 2015 October, 26]. Available from: http://www.qu.edu.qa/pharmacy/components/upcoming_events_material/Qatar_National_Vision_2030.pdf.

74. Hambleton RK, Merenda PF, Spielberger CD. Adapting educational and psychological tests for cross-cultural assessment: Psychology Press; 2004.

75. Malda M, FJR vdV, Srinivasan K, Transler C, Sukumar P, Rao K. Adapting a cognitive test for a different culture: An illustration of qualitative procedures. Psychology science quarterly. 2008;50(4):451-68.

76. Hambleton RK, Kanjee A. Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological Assessment. 1995;11(3):147.

77. Geisinger KF. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment. 1994;6(4):304-12.

78. Lee Y-M, Ahn D-S. The OSCE: a new challenge to the evaluation system in Korea. Medical teacher. 2006;28(4):377-9.

79. Selim AA, Ramadan FH, El-Gueneidy MM, Gaafer MM. Using Objective Structured Clinical Examination (OSCE) in undergraduate psychiatric nursing education: is it reliable and valid? Nurse education today. 2012;32(3):283-8.

80. Huang YS, Liu M, Huang CH, Liu KM. Implementation of an OSCE at Kaohsiung Medical University. Kaohsiung J Med Sci. 2007;23(4):161-9.

81. Austin Z, Gregory PA. Evaluating the accuracy of pharmacy students' self-assessment skills. American journal of pharmaceutical education. 2007;71(5).

82. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, Van Der Vleuten C, Van der Velden U. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. European Journal of Dental Education. 2009;13(3):162-71.

83. AACP. Doctor of Pharmacy (Pharm.D.) Degree American Association of Colleges of Pharmacy2016 [cited 2016 18 January]. Available from: http://www.aacp.org/resources/student/pharmacyforyou/Documents/PharmDCurriculum.htm.

84. Boursicot KAM, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Medical education. 2007;41(11):1024-31.

85. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Advances in Health Sciences Education. 2006;11(2):115-22.

86. Belfast QsU. Standard setting 2012 [cited 2016 18 January]. Available from: http://www.med.qub.ac.uk/osce/background_Standard.html.

87. Hoskin T. Parametric and nonparametric: Demystifying the terms. Mayo Clinic CTSA BERD Resource Retrieved from http://www mayo edu/mayo-edudocs/center-for-translational-science-activities-documents/berd-5-6 pdf. 2014.

88. Lukas RV, Adesoye T, Smith S, Blood A, Brorson JR. Student assessment by objective structured examination in a neurology clerkship. Neurology. 2012;79(7):681-5.

89. Mittal MK, Dhuper S, Siva C, Fresen JL, Petruc M, Velázquez CR. Assessment of email communication skills of rheumatology fellows: a pilot study. Journal of the American Medical Informatics Association. 2010;17(6):702-6.

90. Reimann C, Filzmoser P. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental geology. 2000;39(9):1001-14.

91. Elliott AC, Woodward WA. Statistical analysis quick reference guidebook: With SPSS examples: Sage; 2007.

92. Kim H-Y. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative dentistry & endodontics. 2013;38(1):52-4.

93. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A psychometric toolbox for testing validity and reliability. Journal of Nursing scholarship. 2007;39(2):155-64.

94. Shultz KS, Whitney DJ, Zickar MJ. Measurement theory in action: Case studies and exercises: Routledge; 2013.

95. Remmen R, Scherpbier A, Denekens J, Derese A, Hermann I, Hoogenboom R, et al. Correlation of a written test of skills and a performance based test: a study in two traditional medical schools. Medical teacher. 2001;23(1):29-32.

96. Sedgwick P. External and internal validity in clinical trials. Bmj. 2012;344.

97. Haynes SN, Richard D, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. Psychological assessment. 1995;7(3):238.

98. Gwet KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters: Advanced Analytics, LLC; 2014.

99. Williams GW. Comparing the Joint Agreement of Several Raters with Another Rater. Biometrics. 1976;32(3):619-27.

100. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276-82.

101. Jonsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review. 2007;2(2):130-44.

102. Varkey P, Natt N, Lesnick T, Downing S, Yudkowsky R. Validity Evidence for an OSCE to Assess Competency in Systems-Based Practice and Practice-Based Learning and Improvement: A Preliminary Investigation. Academic Medicine. 2008;83(8):775-80.

103. Henson RK. Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. Measurement and evaluation in counseling and development. 2001;34(3):177.

104. Monaghan MS, Vanderbush RE, Allen RM, Heard JK, Cantrell M, Randall J. Standardized patient use outside of academic medicine: Opportunities for collaboration between medicine and pharmacy. Teaching and learning in medicine. 1998;10(3):178-82.

105. George D, Mallery M. Using SPSS for Windows step by step: a simple guide and reference. Boston, MA: Allyn y Bacon[Links]. 2003.

106. Gliem JA, Gliem RR, editors. Calculating, interpreting, and reporting Cronbach’s alpha reliability coefficient for Likert-type scales2003: Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.

107. Rizzo AA, Kim GJ. A SWOT analysis of the field of virtual reality rehabilitation and therapy. Presence. 2005;14(2):119-46.

108. Merriam SB. Qualitative research: A guide to design and implementation: John Wiley & Sons; 2014.

109. Curry LA, Nembhard IM, Bradley EH. Qualitative and Mixed Methods Provide Unique Contributions to Outcomes Research. Circulation. 2009;119(10):1442-52.

110. Patton MQ. Qualitative evaluation and research methods: SAGE Publications, inc; 1990.

111. Coyne IT. Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries? Journal of Advanced Nursing. 1997;26(3):623-30.

112. Morgan DL. The focus group guidebook: Sage publications; 1997.

113. Sim J. Collecting and analysing qualitative data: issues raised by the focus group. Journal of Advanced Nursing. 1998;28(2):345-52.

114. Kajornboon AB. Using interviews as research instruments. E-journal for Research Teachers. 2005;2(1).

115. Bailey J. First steps in qualitative data analysis: transcribing. Family Practice. 2008;25(2):127-31.

116. Bentley E. oTranscribe BETA [cited 2015 June 25th]. Available from: http://otranscribe.com/.

117. Hsieh H-F, Shannon SE. Three Approaches to Qualitative Content Analysis. Qualitative Health Research. 2005;15(9):1277-88.

118. Elo S, Kyngas H. The qualitative content analysis process. J Adv Nurs. 2008;62(1):107-15.

119. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, Van Der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Medical education. 2003;37(2):132-9.

120. Martin IG, Jolly B. Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. Medical education. 2002;36(5):418-25.

121. Tamblyn R, Abrahamowicz M, Dauphinee W, et al. Association between licensure examination scores and practice in primary care. JAMA. 2002;288(23):3019-26.

122. Tamblyn R, Abrahamowicz M, Brailovsky C, et al. Association between licensing examination scores and resource use and quality of care in primary care practice. JAMA. 1998;280(11):989-96.

123. Vallevand A, Violato C. A predictive and construct validity study of a high-stakes objective clinical examination for assessing the clinical competence of International Medical Graduates. Teaching and learning in medicine. 2012;24(2):168-76.

124. Remmen AS, Joke Denekens, Anselm Derese, Ingeborg Hermann, Ron Hoogenboom, Cees van der Vleuten, Paul van Royen, Leo Bossaert, Roy. Correlation of a written test of skills and a performance based test: a study in two traditional medical schools. Medical teacher. 2001;23(1):29-32.

125. McLaughlin JE, Khanova J, Scolaro K, Rodgers PT, Cox WC. Limited Predictive Utility of Admissions Scores and Objective Structured Clinical Examinations for APPE Performance. American journal of pharmaceutical education. 2015;79(6):84.

126. Barman A. Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singapore. 2005;34(8):478-82.

127. Kirton SB, Kravitz L. Objective Structured Clinical Examinations (OSCEs) compared with traditional assessment methods. American journal of pharmaceutical education. 2011;75(6):111.

128. Munoz LQ, O'Byrne C, Pugsley J, Austin Z. Reliability, validity, and generalizability of an objective structured clinical examination (OSCE) for assessment of entry-to-practice in pharmacy. Pharmacy Education. 2005(5):33-43.

129. Gormley G. Summative OSCEs in undergraduate medical education. The Ulster Medical Journal. 2011;80(3):127-32.

130. Hamdy H, Prasad K, Williams R, Salih FA. Reliability and validity of the direct observation clinical encounter examination (DOCEE). Medical education. 2003;37(3):205-12.

131. Eldarir SA, el Hamid NAA. Objective Structured Clinical Evaluation (OSCE) versus Traditional Clinical Students Achievement at Maternity Nursing: A Comparative Approach. IOSR Journal of Dental and Medical Sciences. 2013;4(3):63-8.

132. Troncon LEdA. Clinical skills assessment: limitations to the introduction of an "OSCE" (Objective Structured Clinical Examination) in a traditional Brazilian medical school. Sao Paulo Medical Journal. 2004;122:12-7.

133. Taylor CA, Green KE. OSCE Feedback: A Randomized Trial of Effectiveness, Cost-Effectiveness and Student Satisfaction. Creative Education. 2013;4(06):9.

134. Brand HS, Schoonheim-Klein M. Is the OSCE more stressful? Examination anxiety and its consequences in different assessment methods in dental education. European Journal of Dental Education. 2009;13(3):147-53.

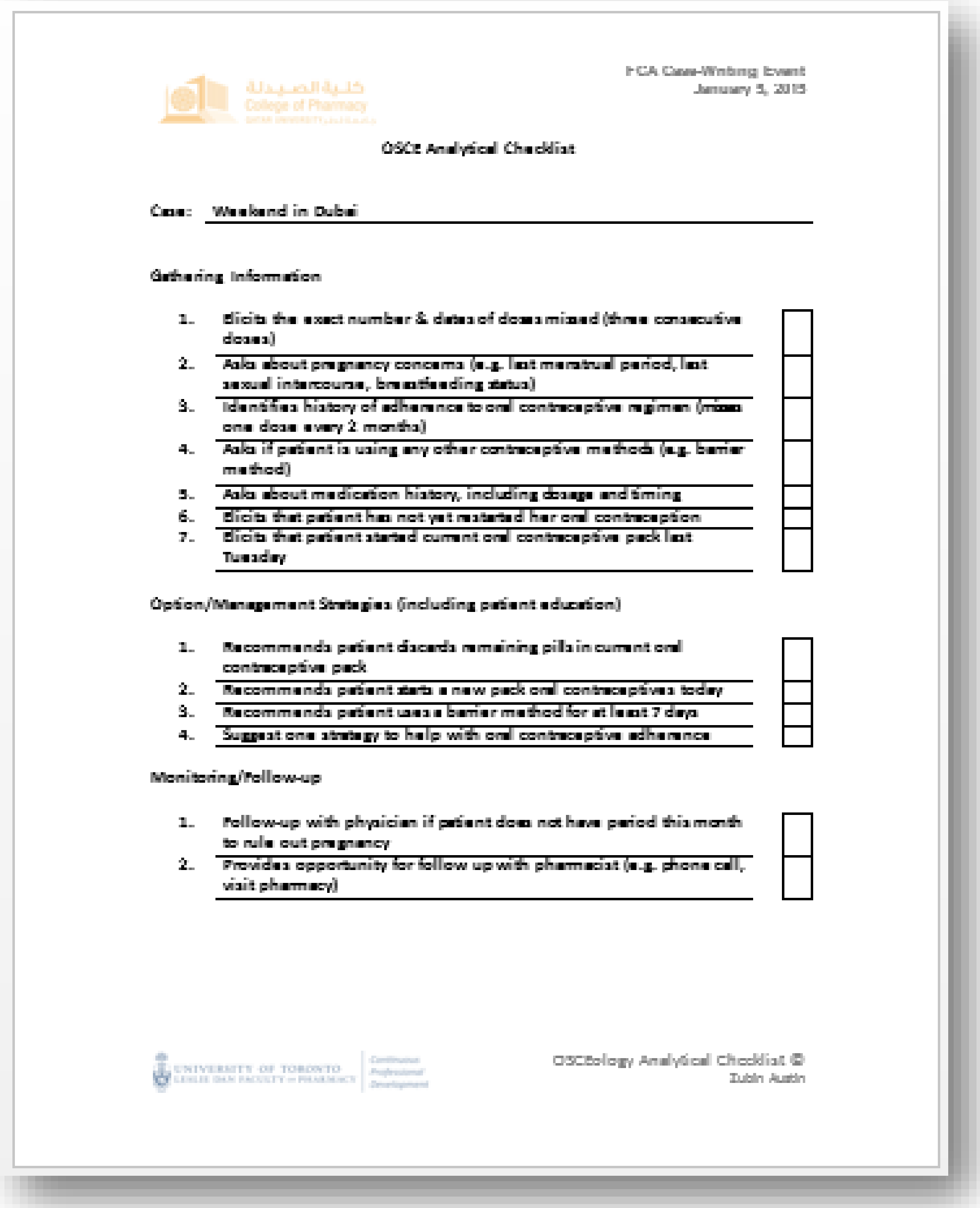
135. Awaisu A, Mohamed MH, Al-Efan QA. Perception of pharmacy students in Malaysia on the use of objective structured clinical examinations to evaluate competence. American journal of pharmaceutical education. 2007;71(6):118.

136. El-Nemer A, Kandeel N. Using OSCE as an assessment tool for clinical skills: nursing students' feedback. Australian Journal of basic and Applied sciences. 2009;3(3):2465-72.

137. Towers DN, Allen JJB. A Better Estimate of the Internal Consistency Reliability of Frontal EEG Asymmetry Scores. Psychophysiology. 2009;46(1):132-42.

138. Collective TD-BR. Design-based research: An emerging paradigm for educational inquiry. Educational Researcher. 2003:5-8.

Appendix A: A Template of analytical checklist instrument used in evaluation of students' performance in an OSCE station



كلية الصيدلة
College of Pharmacy
QATAR UNIVERSITY

FCA Case-Writing Event
January 5, 2015

OSCE Analytical Checklist

Case: Weekend in Dubai

Gathering Information

1. Elicits the exact number & dates of doses missed (three consecutive doses) □
2. Asks about pregnancy concerns (e.g. last menstrual period, last sexual intercourse, breastfeeding status) □
3. Identifies history of adherence to oral contraceptive regimen (misses one dose every 2 months) □
4. Asks if patient is using any other contraceptive methods (e.g. barrier method) □
5. Asks about medication history, including dosage and timing □
6. Elicits that patient has not yet restarted her oral contraception □
7. Elicits that patient started current oral contraceptive pack last Tuesday □

Option/Management Strategies (including patient education)

1. Recommends patient discards remaining pills in current oral contraceptive pack □
2. Recommends patient starts a new pack oral contraceptives today □
3. Recommends patient uses a barrier method for at least 7 days □
4. Suggest one strategy to help with oral contraceptive adherence □

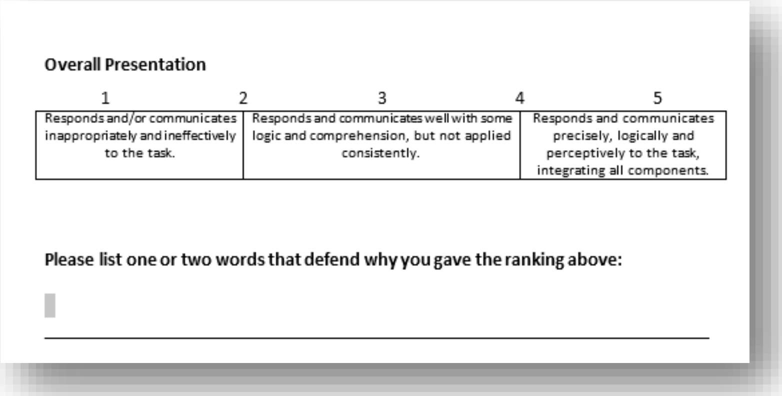
Monitoring/Follow-up

1. Follow-up with physician if patient does not have period this month to rule out pregnancy □
2. Provides opportunity for follow up with pharmacist (e.g. phone call, visit pharmacy) □

UNIVERSITY OF TORONTO
LESLIE DAN FACULTY of PHARMACY
Continuing Professional Development

OSCEology Analytical Checklist ©
Zubin Austin

Appendix B: A template of global scoring instrument used in evaluation of students' performance in an OSCE station



Overall Presentation

1 2 3 4 5

Responds and/or communicates inappropriately and ineffectively to the task.	Responds and communicates well with some logic and comprehension, but not applied consistently.	Responds and communicates precisely, logically and perceptively to the task, integrating all components.

Please list one or two words that defend why you gave the ranking above:

Appendix C: Table showing the blue print of the OSCE stations describing their topics, focus, and complexity.

Station	**Disease**	**Focus**	**Complexity**
1	Cold sore	Counseling	Simple problem Simple patient
2	Bacterial conjunctivitis	Education	Simple problem Simple patient
3	Breast cancer	Calculation	Complex problem Simple patient
4	Food Poisoning	Referral	Complex problem Simple patient
5	Cardiology	Adverse reaction management	Simple problem Simple patient
6	Pharmacy Management	Staff supervision	Complex problem Simple patient
7	Diabetes	Education	Simple patient Complex problem
8	Mental Health	Problem recognition and communication	Complex problem Complex patient
9	Tuberculosis	Patient assessment	Complex problem Simple patient
10	Contraception	Drug information	Simple problem Simple patient

Appendix D: Table describes internal consistency

Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent
0.9 > α ≥ 0.8	Good
0.8 > α ≥ 0.7	Acceptable
0.7 > α ≥ 0.6	Questionable
0.6 > α ≥ 0.5	Poor
0.5 > α	Unacceptable