PSYCHOMETRIC ANALYSIS AS A QUALITY ASSURANCE SYSTEM IN OSCEs IN A RESOURCE LIMITED INSTITUTION

by

DR A.O. OGAH

Thesis submitted in fulfilment of the requirements for the degree of Philosophiae Doctor in Health Professions Education
Ph.D. HPE

in the

Division Health Sciences Education,
Faculty of Health Sciences
at the University of the Free State

July 2016

Promoter: Dr M.P. Jama
Co-Promoter: Prof. H. Brits

DECLARATION

I hereby declare that the compilation of this dissertation is the result of my own independent investigation. I have endeavoured to use the research sources cited in the text in a responsible way and to give credit to the authors and compilers of the references for the information provided, as necessary. I have also acknowledged those persons who have assisted me in this endeavour. I further declare that this work is submitted for the first time at this University and faculty for the purpose of obtaining a Philosophiae Doctor degree in Health Professions Education and that it has not previously been submitted to any other university or faculty for the purpose of obtaining a degree. I also declare that all information provided by study participants will be treated with the necessary confidentiality.

July 2016

DR A.O. OGAH **Date**

I hereby cede copyright of this product in favour of the University of the Free State.

July 2016

DR A.O. OGAH **Date**

DEDICATION

I dedicate this thesis to JESUS CHRIST, who initiated it, carried it through and has completed it.

ACKNOWLEDGEMENTS

I wish to express my sincere thanks and appreciation to the following persons:

- My Promoter, Dr Mpho Jama, Head: Student Learning and Development, Faculty of Health Sciences, University of the Free State, for her incredible support, mentorship, expert supervision and patience. She has become more like a Mother to me than just a Supervisor.
- My Co-Promoter, Prof. Hanneke Brits, Department of Family Medicine, School of Medicine, University of the Free State, for her support and valuable advice and contributions. She has helped me to focus on one and only one thing, one thing…
- My HPE Programme Coordinator, Dr Johan Bezuidenhout, Head: Division Health Sciences Education, Faculty of Health Sciences, University of the Free State, for his contributions as Coordinator and Facilitator of the HPE programme. He is a very strong driver of the programme and the friend of students because of his humble, generous and kind personality.
- My Consultant Biostatistician, Ms Maryn Viljoen, for her generous support and advice.
- Mrs Elmarié Robberts, Secretary: Division Health Sciences Education, Faculty of Health Sciences, University of the Free State for the formatting and editing of and giving meticulous attention to all technical detail of this thesis.
- Dr Medadi Ssentanda, African Language department, Makerere University, Uganda, for assisting in the edition of this thesis.
- My husband, Architect Anthony Ogah, who offered me unconditional love and support throughout the course of this thesis.
- My daughter and sons, Elizabeth, James, Elijah and Daniel, for their love and support throughout the course of this thesis.
- My Deputy Vice Chancellor, Dr Abanis, DVC: Kampala International University, Dar campus, for his support, encouragement and mentorship.
- Mr Thomas, the Director of ICT, Kampala International University, Dar es Salaam, Tanzania for ensuring I had internet access in my office and for the general maintenance of internet supply to KIU-Dar community. His good work increased my speed in completing this thesis.
- Colleagues and friends, who supported, encouraged and assisted with patience, during the course of the thesis.
- The students and examiners who participated in this study, for allowing me to observe

you during your examinations - without your time and cooperation, this project would not have been possible and last but not the least,

- To my Heavenly Father who gave me the courage to attempt the study, the means, the strength and perseverance to complete it.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND ORIENTATION TO THE STUDY

CHAPTER 2: PSYCHOMETRIC ANALYSIS AS A QUALITY ASSURANCE SYSTEM IN THE OSCEs

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

CHAPTER 4: RESULTS OF THE PSYCHOMETRIC ANALYSIS OF THE OSCEs

CHAPTER 5: INTERPRETATION AND DISCUSSION: PSYCHOMETRIC ANALYSIS OF THE OSCEs

CHAPTER 6: CONCLUSION, RECOMMENDATIONS AND LIMITATIONS OF THE STUDY

APPENDICES:

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACGME:	**Accreditation Council of Graduate Medical Education**
AMEE:	**Association of Medical Educators in Europe**
ANOVA:	**Analysis of Variance**
BLR:	**Borderline Regression Methods**
CTT:	**Classical Test Theory**
d:	**Discrimination Index**
ENT:	**Ear, Nose and Throat**
HDI:	**High Development Index**
IDI:	**Item Difficulty Index**
IMG:	**International Medical Graduates**
IRT:	**Item Response Theory**
KIU:	**Kampala International University**
KIU-Dar:	**Kampala International University-Dar Campus**
KR-20:	**Kuder–Richardson 20**
LDI:	**Low Development Index**
MB.,Ch.B:	**Medicinae Baccalaureus, Baccalaureus Chirurgiae**
OSCE:	**Objective Structured Clinical Examination**
QAS:	**Quality Assurance System**
r:	**Pearson Correlation**
r^2:	**Importance of the Station**
SD:	**Standard Deviation**
SEM:	**Standard Error of the Mean**
SPs:	**Standardised/Simulated patients**
SPSS:	**Statistical Package for Social Sciences**
UK:	**United Kingdom**

DEFINITIONS OF TERMS

The definitions applied in this study are adopted from various authors as referenced hereunder.

OSCE: (Objective Structured Clinical Examination) is an assessment tool based on the principles of objectivity and standardisation, in which the candidates move through a series of time-limited stations in a circuit for the purposes of assessment of professional performance in a simulated environment. At each station, candidates are assessed and marked against standardised scoring rubrics by trained assessors (Kamran, Sankaranarayanan, Kathryn & Piyush 2013a:e1438).

Reliability: The reproducibility of assessment scores over time or occasions (Downing 2004:1007).

Validity: The degree of meaningfulness for any interpretation of a test score (Downing, 2005:350).

Psychometric or item analysis: It is a quantitative, statistical or mathematical method of analysing human behaviour or attributes as represented in figures (Tavakol & Dennick 2012:e162).

Standardised Patients: These are OSCE role-play actors used to simulate patients according to a pre-determined script (Adamo 2003:263).

Limited Resource Institution: Restricted amounts of inputs required by an Institution such as motivated staff, finances, production facilities and raw materials (Business dictionary 2014:1 of 1).

Low Development Index Countries: according to the United Nations, these are countries with a lower living standard, underdeveloped industrial base and low human development index relative to other countries. The human indices used are lower life expectancy, less education and less income (Sullivan & Steven 2014:1 of 10).

High Development Index: according to the United Nations, these are sovereign countries with highly developed economy and advanced technological infrastructure relative to other countries (Investopaedia 2010:1 of 11).

SUMMARY

Key terms: OSCE, Psychometric analysis, Quality Assurance System, Resource Limited Medical Schools, improving assessments, examination policy, improving clinical training, improving quality of clinical graduates, patient safety, post-examination evaluation.

A comprehensive study was carried out with a view to develop a guideline for psychometric analysis and recommend for its incorporation into the Quality Assurance examination policy of medical schools. The study was motivated by a gap that exists in the knowledge and skills of psychometric methods at KIU-Dar, the UFS, but also in Sub- Saharan Africa and the rest of Africa. Psychometrics involves statistics and mathematics, which is a nightmare for many practitioners. To bridge the gap, the researcher compiled a very simple guideline for psychometric analysis. The guideline discussed is user-friendly and requires simple easily accessible tools such as SPSS and Microsoft Excel. Moreover, a new psychometric programme for the rapid analysis of the OSCE data, developed and published by Tavakol and Doody early this year, was introduced in this study with recommendations for medical schools to purchase it for use in their medical education units to quality assure their examinations. By developing the strategy, the identified gap was bridged, in that it can aid in the training and encouragement of academic staffs to use the psychometric tools, skilfully on their subject examinations to improve assessments and training.

The study was carried out in a resource limited medical school, at the Kampala International University, Dar es Salaam campus, to objectively measure the quality of the OSCEs in order to improve and harmonize the quality of our assessments, clinical training, quality of medical graduates and consequently patient care. The exit OSCE was conducted in four clinical departments (OBGYN, Paediatrics, Medicine and Surgery) in July 2015. The examiners were a mixture of consultants, specialists and medical officers. The research methods comprised of literature reviews and observations using checklists during the OSCEs. The examiners collected the scores (data) by means of checklists. A total of 27 graduating clinical students were assessed by 20 examiners in 82 OSCE stations across all the departments.

The key findings in this study were as follows. The examiners were too few, while the OSCE stations were too many in 3 of the departments. OBGYN had too few OSCE stations. Of note is the very limited human resource in OBGYN, as all the internal examiners were part-

time staff of the university. The examiner: station ratio was 1:4. The clinical stations were manned by one examiner each and the rest of the stations (50-75%) were in written format. The examiners were not very conversant with OSCE, they were not aware of global scoring, standard setting and psychometric analysis. Currently, our medical school uses a fixed university pass mark of 50% and the grading system is based on the raw scores. In addition, there was no blueprint and standardised patient for the OSCE. All the patients that were used were real, which limited the researcher from obtaining their ratings of the students' performances as this was not permitted by the hospital administration. Also, the researcher could not obtain the examiners' global scores. The Ministry of Health checklist that was used by the examiners did not capture global scoring nor standardised patients' ratings. Since it was practically difficult to have an ideal examination setting, the raw OSCE scores were likely to be abnormally distributed with possibly outliers as we found in this study.

The findings of the study inform helpful recommendations pertaining to accurate perception of the students' performances and ability to make valid comparison with other assessments anywhere. The study suggests that it is better to convert the raw scores into Z-scores and the university grading systems should be built on Z-scores. As was also observed in the analysis, the letter grades based on the Z-scores corresponded to different raw score ranges in each station and subject. Moreover, to make an accurate pass-fail judgement on students' performances, it is better to use the 'gold standard' method of setting the pass mark especially in a situation where the examination setting is not ideal. The borderline regression method is the recommended method. However, where there are no global scores, the borderline method, as was demonstrated in this study can be used to determine the pass mark for these examinations.

The variability of the OSCE scores was generally high, which is not desirable for a criterion referenced test. Also, the variability in an ideal OSCE should only come from the different abilities of the students. However, in this study, most of the variability in the students' scores was contributed by the examiners and their interactions with the students as noted in the ANOVA and the G-studies. The overall G-Coefficient was 0.75, which is comparable with other studies in the developed countries. However, the G-coefficient in each subject were lower and weakest in OBGYN. The station analysis in this study showed that the Internal Medicine OSCE was the best. The other subjects especially OBGYN had poor discriminating power and difficulty index for a criterion referenced assessment. The internal consistency and stability of the OSCEs in this resource constrained institution, based on

Cronbach alpha, were low (below 0.25). Several Hawks and Doves were identified in the study. Recommendations in this regard were made. The sound research approach and methodology ensured quality, reliability and validity. The completed research can form the basis for a further research undertaking.

OPSOMMING

Sleutelterme: OSCE, psigometriese ontleding, gehalteversekeringstelsel, mediese skole met hulpbronbeperkings, verbetering van assessering, eksamenbeleid, verbetering van kliniese opleiding, verbetering van die gehalte van kliniese gegradueerdes, pasiëntveiligheid, ander fakulteite.

'n Omvattende studie is uitgevoer met die doel om 'n riglyn te ontwikkel vir psigometriese ontleding en om dit in die gehalteversekeringseksamenbeleid van elke mediese skool in te sluit. Dit sal begin by die mediese skole van die Universiteit van die Vrystaat en Kampala International University. Die riglyn, as dit gereeld gebruik word, sal die gehalte van die OSCEs en ander belangrike universiteitseksamens konsekwent en objektief meet, met die doel om die gehalte van assesserings, kliniese opleiding, die gehalte van mediese gegradueerdes en, gevolglik, pasiëntsorg, te verbeter en in ooreenstemming te bring.

Assessering is die kern van elke opleidingsinstelling. Assessering dryf leer. Deur assessering kan die hoeveelheid leer wat die student verwerf het, gemeet word, en kan die programme geëvalueer word. Geen assessering is ooit presies dieselfde as 'n ander nie. Die toepassing van OSCE, 'n soort kliniese assessering wat hulpbronintensief is, kan gekompromitteer word ten opsigte van gehalte, veral by instellings met beperkte hulpbronne. Die gehalteversekeringsmetodiek wat tans by die meeste mediese skole ingestel is om die gehalte van assessering te monitor, gebruik mense as beoordelaars, wat subjektief, bevooroordeeld en veranderlik kan wees. Psigometriese ontleding bied 'n stabiele, objektiewe en goedkoop manier om die gehalte van OSCEs en alle ander tipes assessering konsekwent te meet en te verbeter.

Daar bestaan reeds heelwat dokumentasie en publikasies oor die psigometriese waarde van OSCEs soos dit by mediese skole in ontwikkelde lande toegepas word, maar selfs daar is dit nog nie vir algemene gebruik in universiteitsbeleid ingesluit nie. Gevolglik lewer elke universiteit mediese gegradueerdes met verskillende vaardigheidsvlakke. Baie min is al gepubliseer oor die ware toedrag van sake van OSCEs by mediese skole in Afrika suid van die Sahara wat hulpbronbeperkings ervaar. Dit kan risiko inhou vir pasiëntveiligheid, veral in lande wat nie nasionale kwalifiserende eksamens het wat die gehalte van assesserings en mediese gegradueerdes wat gesertifiseer en geregistreer word vir die praktyk, harmonieer en reguleer nie. Daarom moet psigometriese ontleding ten volle by die gehalteversekeringseksamenbeleid van elke mediese skool geïntegreer word, ten einde die gehalte van assesserings, opleiding, mediese gegradueerdes en, gevolglik, pasiëntsorg, te harmonieer en te verbeter.

Hierdie studie het psigometriese metodes beskryf en die aanwending daarvan op 'n eenvoudige wyse geïllustreer ten einde die psigometriese eienskappe van die finale OSCE, soos dit by 'n mediese skool met beperkte hulpbronne in oos-Afrika (Dar es Salaam) toegepas is, te meet, interpreteer en bespreek. Die mediese skool wat vir hierdie studie gebruik is, is 'n private instelling wat ten tye van die ondersoek vier jaar oud was en nie sy eie opleidingshospitaal het nie (dit was in aanbou). Die mediese skool is betrokke by 'n personeel-student uitruil-opleidingsprogram, wat personeel en studente wat op verskillende opleidingsvlakke is by die Dar-kampus, in staat stel om uit te ruil met ander studente van susterkampusse in Uganda en Nairobi. Die Dar-kampus se kliniese studente roteer na nabygeleë verwysingshospitale, wat gestruktureer is om basiese pasiëntsorg te verskaf en wat by die universiteit geaffilieer is. Daarom word die finale OSCE by een van die streeks-verwysingshospitale in Dar es Salaam uitgevoer.

Die navorsingsmetodes het bestaan uit 'n literatuurstudie en waarnemings deur die navorser, wat 'n oorsiglys gedurende die OSCEs gebruik het. Die literatuuroorsig het agtergrond verskaf vir 'n konseptuele raamwerk en het die probleem teen verbandhoudende teorie en navorsing gekontekstualiseer. Die eksaminatore het die tellings (data) deur middel van oorsiglyste versamel. 'n Totaal van 27 graduerende kliniese studente is geëksamineer.

Die bevinding van hierdie studie is soos volg. Die finale OSCE is in vier kliniese departmente uitgevoer: Obstetrie en Ginekologie, Pediatrie, Geneeskunde en Chirurgie. Die eksaminatore was 'n mengsel van konsultante, spesialiste en mediese beamptes. By drie departemente was daar te min eksaminatore en te veel OSCE stasies, en te min stasies by Obstetrie en Ginekologie. Die eksaminator:stasie verhouding was 1:4. Die kliniese stasies is elk deur een eksaminator beman, en die res van die stasies (75%) het geskrewe response vereis. Die eksaminatore was nie juis bekend met OSCE nie: Hulle was nie bewus van oorhoofse puntetoekenning, standaardbepaling of psigometriese ontleding nie. Die mediese skool gebruik tans 'n vaste universiteitslaagsyfer van 50%, en die stelsel van puntetoekenning is op die rou tellings gebaseer.

Daar bestaan geen bloudruk of gesimuleerde pasiënt vir die OSCE nie. Al die pasiënte wat gebruik is, was ware pasiënte, wat die navorsing beperk het, aangesien die hospitaaladministrasie nie gewillig was om studente se prestasiegraderings te verskaf nie, en die navorsing kon ook nie toegang kry tot die eksaminatore se oorhoofse tellings nie. Die oorsiglys van die Ministerie van Gesondheid wat die eksaminatore gebruik het, het nie plek gehad vir oorhoofse tellings of vir gesimuleerde pasiënttellings nie. Omdat dit prakties moeilik is om 'n ideale eksamensituasie te hê, sal die rou OSCE tellings waarskynlk abnormaal versprei wees, met moonlike uitskieters, en dit is deur dié studie bevestig.

Daarom, om 'n akkurate persepsie van die studente se prestasies te verkry, en om dit moontlik te maak om 'n geldige vergelyking met ander assesserings van enige ander plek, te doen, is dit beter om die rou tellings in Z-tellings om te skakel, en om universiteite se puntestelsels op Z-tellings te baseer. Soos in die ontleding waargeneem is, het die letterpunte wat op die Z-tellings gebaseer is met die verskillende routellingreekse by elke stasie, en vir elke vak, ooreengestem. Verder, om op grond van studente se prestasies 'n akkurate slaag-druip oordeel te maak, is dit beter om die "goudstandaard" metode vir die bepaling van die slaagsyfer te gebruik, veral in 'n situasie waar die eksamensituasie nie ideaal is nie. Die grensgeval-regressiemetode is egter die metode wat aanbeveel word. Waar daar geen oorhoofste punte is nie, kan die grensgevalmetode, soos deur hierdie studie getoon, gebruk word om die slaagpunt vir elke eksamen te bereken.

Die veranderlikheid van OSCE punte was oor die algemeen hoog, wat nie gunstig is vir 'n kriteriumverwysingstoets nie. Verder moet die bron van die veranderlikheid van 'n ideale OSCE net die verskillende vermoëns van die studente wees. In hierdie studie is veranderlikheid in die studente se punte ook deur die eksaminatore en hulle interaksie met die studente veroorsaak, soos deur die ANOVA en G-studies getoon. In hierdie studie het 'n ontleding van die stasies getoon dat die stasie by Interne Geneeskunde die beste een was. Die ander vakke het swak gediskrimineer en die moeilikheidsindeks vir 'n kriteriumverwysde assessering was swak. Die interne konsistensie en stabiliteit van die OSCEs in hierdie instelling, met sy beperkte hulpbronne, was, volgens die Cronbach *alfa*, laag (onder 0.25). Verskeie valke en duiwe is in die studie geïdentifiseer.

Die studie het onstaan uit die besef dat daar 'n gaping bestaan tussen die kennis en vaardighede van psigometriese metodes by Kampala International University en die Universiteit van die Vrystaat, maar ook in Afrika Suid van die Sahara en die res van Afrika

oor die algemeen. Psigometrie behels Statistiek en Wiskunde, wat vir baie praktisyns 'n nagmerrie verteenwoordig. Om die gaping te oorbrug, het die navorser 'n baie eenvoudige riglyn vir psigometriese ontleding saamgestel en maak sy voorstelle vir die integrasie van psigometriese ontleding van belangrike eksamens, waaronder OSCEs, as 'n noodsaaklike komponent van die huidige gehalteversekeringstelsel. Die doel hiermee is om objektief en konsekwent die gehalte van eksamens te meet, en om die gehalte van belangrike eksamens, opleiding, leer, mediese gegradueerdes en pasiëntsorg te verbeter.

Die gebruikersvriendelike psigometriese riglyn word bespreek, asook die eenvoudige hulpmiddels wat beskikbaar is vir gebruik, met die doel om te verseker dat elke eksaminator kennis en vaardighede met betrekking tot psigometriese ontleding bekom, en in staat sal wees om dit gereeld in eksamens by hulle instellings toe te pas ten einde assessering te verbeter. Verder het hierdie studie 'n nuwe psigometriese program vir die vinnige ontleding van OSCE, wat deur Tavakol en Doody (2016) ontwikkel is, bekend gestel, en word mediese skole aanbeveel om dit aan te koop en in hulle mediese opleidingseenheide te gebruik vir gehalteversekering van hulle eksamens. Deur die strategie te ontwikkel, is die gaping wat geïdentifiseer is, oorbrug en kan dit 'n bydrae maak tot die opleiding en aansporing van akademiese personeel om hulle vaardighede met die gebruik psigometriese hulpbronne te verbeter, en daardeur ook assessering en opleiding te verbeter. Aanbevelings word in hierdie verband gemaak. Die grondige navorsingsbenadering en -metodologie het die gehalte, betroubaarheid en geldigheid van die studie verseker. Die voltooide navorsing kan die grondslag vorm van verdere navorsingsondernemings.

PSYCHOMETRIC ANALYSIS AS A QUALITY ASSURANCE SYSTEM IN OSCE IN A RESOURCE LIMITED INSTITUTION

CHAPTER 1

INTRODUCTION AND ORIENTATION TO THE STUDY

1.1 INTRODUCTION

In this project, the researcher carried out an in-depth study to illustrate the use of psychometric methods to analyse and interpret Objective Structured Clinical Examination (OSCE) scores with a view to monitor and improve assessment and therefore ensure quality in a resource limited institution.

The main aim in this study was to illustrate the use of psychometric methods to analyse and interpret raw OSCE scores of the graduating clinical medicine students' final assessment. The OSCEs were carried out in four clinical departments, namely, Obstetrics and gynaecology, Paediatrics, Internal Medicine and Surgery at the School of Health Sciences, Kampala International University (KIU-DAR-DAR), Dar es Salaam campus (hereafter referred to as KIU-Dar) in Tanzania. With this in mind, the study informs the reader of the available psychometric methods, their application in the analysis and interpretation of post-examination live OSCE scores from a medical school with limited resources, thus ensuring quality in assessment.

Louise (2002:502) documented that the current chief concern of medical educators globally is the promotion of professional behaviour. A critical component of this concern involves assessment. The author emphasised that the progress of learners in becoming and being professional and the success of programs that promote professionalism can be measured and ascertained through assessment. Hence, medical educators are saddled with the responsibility of assessing both the learners and programs. Assessment being a driver of learning is a vital part of the training of all students. Therefore, the central role of assessment in promoting professionalism requires an examination of state of the art (Louise 2002:502).

The characteristics of assessment tools identified by Van der Vleuten and Schuwirth (2005:310) include the following aspects: validity, reliability, educational impact, feasibility, acceptability and cost. The first aspect, validity, refers to the extent to which the tool measures what it is supposed to measure. The second aspect, reliability, is the tool's ability to yield consistent results each time it is used. The third aspect, which is educational impact, means the effect of the tool on teaching and learning. The fourth aspect, feasibility, considers the practicability of implementing the assessment tool in context. The fifth aspect, acceptability, refers to the readiness of all the stakeholders (staff, students and others) to adopt a particular assessment tool as a method of student evaluation. The last aspect, which is cost, involves the amount of financial, personnel, time and efforts needed to implement the assessment method correctly. Schwartz (2011:2) further stresses the importance of choosing the most appropriate assessment, continually monitoring and improving the quality of the tests that will determine the competence of future healthcare professionals.

According to Crossley, Davies, Humphris and Jolly (2002:972), the reliability of professional performance assessments, such as the OSCE, is threatened by the complexity of the professional behaviour which are intangible and varies markedly from setting to setting and from case to case.

Several authors, such as Turner and Dankoski (2008:574) and Adamo (2003:263) have documented the global transformation of students' clinical assessment methods from a format of written tests of knowledge and traditional long-short case examinations to the OSCE, over the past 50 years. For example, in the authors' reports, the medical councils of Canada, Japan, and Korea employ OSCE in their licensing examination. In addition, the National Board of Medical Examiners incorporates the use of OSCE into the United States Medical Licensing Step 3 Examinations and nearly all medical schools in the US and the UK (United Kingdom) have reported the use of OSCE in their regular evaluations. These clinical assessment tools are expected to measure the knowledge, skills and behaviour of the medical student. Also, according to these authors, no one tool can assess all these learning domains effectively. Hence, the authors recommended a combination of assessment tools (test battery approach) and OSCE plays a very important role in this approach of evaluating students' performance in the clinical setting.

Kamran, Sankaranarayanan, Kathryn and Piyush (2013a:e1440) describe an OSCE as an assessment tool that is based on the principles of objectivity and standardization. The objectivity of the OSCE depends on the presentation of the same test to all the candidates. Harden (1988:19) was the first medical educator to demonstrate an OSCE set up, whereby, the candidates move through a series of time-limited stations in a circuit for the purposes of assessment of professional performance in a simulated environment. The design of this assessment could be single, bilateral or multiple parallel circuit/s of similar stations. At each station, candidates are assessed with different specific tasks and marked against standardised or structured scoring checklists by trained assessors. According to Boursicot, Roberts and Pell (2007:1025), an OSCE is at the 'shows how' level of clinical examination in Miller's pyramid of competencies. The OSCE, as a clinical examination assesses clinical skills in history taking, physical examination technique, practicals, procedural skills, patient care and communication skills.

Kamran *et al*. (2013b:e1450) describe the major components of OSCE design to include: blueprinting, station, tasks and checklist development, recruitment and training of examiners, standardized patients and helpers, administration of the OSCE, scoring, marking and standardized testing. From the researcher's opinion, these components of the OSCE are largely unfamiliar to most medical schools in resource limited areas and this may compromise the quality of our OSCE. There is therefore a need to regularly evaluate our OSCEs. Evaluation of assessments is not common practice in many institutions. Unpublished observations have reported the successes of institutions in high development index (HDI) countries with adequate resources in their practice of the OSCE, but the attitude towards OSCE might be different in resource limited institutions.

Some researchers such as Mohammed (2008:1804); Roberts, Newble, Jolly, Reed and Hampton (2006:535) and Boursicot *et al*. (2007:1024) have published articles on the OSCE in developed institutions. However, the subject of the OSCE is still relatively new in institutions with limited resources and in low development index (LDI) countries such as Tanzania, hence the paucity of published research on the OSCE experience in these countries. The struggle with implementing the standard OSCE is still a challenge in some regions of Africa because of its cost, and this might compromise the quality of assessments and ultimately graduates and patient care.

Tavakol and Dennick (2012:e162) assert that the assessments of clinical competence are subject to many potential sources of error especially biases from the rater judgment. Therefore, efforts to identify and reduce the measurement errors and biases due to poor test design or variation in test items, judges, patients or examination procedures will produce standard, valid and reliable assessments.

In Turner and Dankoski's (2008:575) opinion, there is a direct relationship between the quality of assessment methods and processes and the quality of the teaching and learning process in any form of educational activity. Hence, undergraduate and postgraduate medical examination data need to be evaluated, in order to understand, monitor, control and improve the quality of assessments. Regarding the OSCE, the authors argued that, contrary to the general acceptance of the OSCE, there have been recent concerns over the heavy reliance on this particular format above other assessment methods in several medical schools. More so, several critiques, such as Al-Naami, El-Tinay, Khairy, Mofti and Anjum (2011:300) and Gupta, Dewan and Singh (2010:915) have challenged the psychometric properties of the OSCE versus other traditional methods. Hence, the association of medical educators including AMEE have strongly recommended regular evaluation of the OSCE especially for high stakes examinations such as the exit, promotional and certification assessments in order to improve programmes and produce quality graduates.

In Downing's (2005:351) report, evidence for quality of assessment can be accumulated quantitatively and or qualitatively. Validity can be measured more qualitatively than quantitatively and the reverse is true for the reliability estimate of a test. The focus of this research was on quantitative methods. Downing (2004:1007) explains that the reliability of an instrument is closely associated with its validity and that an instrument cannot be valid unless it is reliable. However, an instrument can be reliable and yet not valid. In other words, from Downing (2004:1007)'s point of view, reliability is a necessary but not sufficient condition for validity while reliability is a major source of validity evidence for all assessments. Hence, reliability and validity add up to dependability of an assessment. Of note is that Van der Vleuten (2000:1217) also points out that validity can take several forms unlike reliability, which can be expressed by a single coefficient. The author grouped the evidences for validity into content-related, criterion-related, and construct-related. Hence, of the five characteristics of an assessment mentioned above, the reliability, followed by the validity are the most important quality properties to be considered when planning a test.

Depending on the purpose of the assessment, Schwartz (2011:12) points out that summative examination need to be more reliable while formative tests need to have more of educational impact than the other attributes of an examination.

Mohammed (2008:1803) and Norcini, Anderson, Bollela, Burch, Joa, Costa, Duvivier, Galbraith, Hays, Kent, Perrott and Roberts (2011:210) describes reliability as the ability of an instrument to give the same results consistently over time or occasions. He sub-categorized reliability into stability and internal consistency. In his explanation, stability means that the examination persistently discriminate students' performance on repetitions. Ideally, stability correlation coefficient should not exceed 0.5. Internal consistency means that scores of an examination in each station would be correlated with scores of all other stations. The internal consistency correlation coefficient should exceed 0.8.

Pell, Fuller, Homer and Roberts (2010:802) define psychometric or item analysis as a quantitative statistical or mathematical method of analysing human behaviour or attributes as represented in figures and is also an important post-examination component of the quality assurance system (QAS) for evaluating assessment tools. The 'psycho' refers to the human behaviour or attribute that is being measured and 'metric' is the statistics of the human behaviour represented in figures. It is objective, defensible and can be readily used to monitor continuous improvement of examinations. In Tavakol and Dennick (2012:e162)'s view, psychometrics can provide diagnostic feedback to improve curricula and teaching strategies. Moreover, with increasing demand for public accountability, routine psychometrics can improve the quality of training and patient care over time (Norcini 2005:880). The other components of the QAS, as documented by Pell *et al.* (2010:803) are standardization, peer review of items/stations, examiner training, external examiner moderation and evaluation. Pell *et al.* (2010:803) recommends that all these components be addressed in every assessment cycle for continuous improvement of an institutional examination system.

In Pell *et al.* (2010:802)'s report, psychometrics provides stable and predictable measures of student performance over time to detect and minimize sources of variation in examination data. However, the knowledge and skills involved in carrying out psychometric analysis as a quality assurance measure are only found in very few learning institutions worldwide (Brannick, Erol-Korkmaz & Prewett 2011:1181).

According to the authors, the few psychometrics articles that have been published are complex and difficult to implement in resource limited settings because the tools are largely inaccessible. Also, the fears of mathematical calculations involved in these methods have kept many instructors away from using these tools for evaluating examinations.

Tavakol and Dennick (2012:e162) describe two general methods of psychometric analysis: the easier older one, which is readily available, being the Classical Test Theory and the newer more complex, less available but more comprehensive method, the Item Response Theory.

In the authors' report (Tavakol & Dennick 2012:e162), the classical test theory (CTT) is concerned with the overall reliability of a test. On the one hand CTT uses descriptive methods to identify sources of measurement error and unreliability in a test, thus minimising them. On the other hand, the Item Response Theory (IRT) using Rasch analysis, D-studies and generalization studies are used to identify sources of measurement error, unreliability as well as interactions between item difficulty and student ability. In this study, the intention was to shed more light on how to apply both theories to the raw scores from the KIU-DAR-DAR OSCE (a resource limited setting) as it is appropriate to use a variety of psychometric measures rather than one so that a more complete picture of the quality of any assessment can be obtained.

Accordingly, this study provides a comprehensive picture of how the OSCE quality may be constructed by using a variety of psychometric measures in a simple manner, and also to consider which characteristics of the OSCE are appropriately judged by which measure(s) and their interpretation.

This study was carried out in a resource limited medical school in Tanzania by measuring the reliability properties (and validity indirectly) of the OSCE using psychometric methods and addressed the issues of: which psychometric techniques to use, how to use them and the interpretation of the results with the ultimate goal of empowering resource limited institutional assessors to adopt and apply them regularly on their examination data for continuous monitoring and improvement. The study adopted a descriptive cross-sectional design utilising all the OSCE results of the final or qualifying examination in the four clinical departments: Paediatrics, obstetrics and gynaecology, Internal Medicine and surgery. In the case of the Classical Test Theory, SPSS statistical software version 17 was used to determine the descriptive statistics (scores distribution, mean, mode, median and standard deviation), measures of variability, passing scores, number of failures, statistical significance, Pearson's correlations, Cronbach's alpha and if item (station) is deleted for internal consistency and stability (Tavakol & Dennick 2011a:54). With Microsoft Excel, the standard error of the mean (SEM), station analysis which include item difficulty index, item discrimination index and statistical significance were computed.

In the case of the Item Response Theory (IRT), the generalizability studies were carried out with SPSS to identify the sources of error and how the various factors of the OSCE influence the students' performances. Quality issues such as at each station level, total student performances per subject and overall scores were addressed. Also, corrective measures to improve on the OSCE quality were discussed based on the results of the psychometric analysis (Pell *et al.* 2010:802).

The few accessible papers on this subject, give one or two metrics as measures of quality in the OSCEs (Pell *et al.* 2010:802), but this study addressed quality issues both at the individual station level and across the complete clinical assessment as a whole, using a battery of psychometric tests obtained from literature review.

1.2 BACKGROUND TO THE STUDY

Crossley *et al.* (2002b:973) note that OSCE measures complex and changing human behaviours, which cannot be reduced to a checklist of observable processes. OSCE also depends heavily on human raters which give subjective judgements about performances and therefore is vulnerable to errors. Moreover, concerns have been raised about the substandard quality of university high-stakes examinations in some regions. Roberts, Newble, Jolly, Reed and Hampton (2006:535) discovered that educationally undesirable assessment methods and practices are still being used by many medical schools partly due to lack of knowledge of the technical aspects of assessment, including the application of robust psychometrics. Hence, some medical schools may be making pass/fail decisions on students' fitness to practice based on assessments that are prone to error. These non-standardized decision-making procedures are disturbing because students who are on the borderline of pass/fail decision making may graduate to become doctors whose clinical performance will give cause for concern. In addition, United Kingdom (Crossley *et al.* 2002b:973) has emphasised that the public should be protected from incompetent doctors.

Furthermore, with an increasing use of criterion-based assessment techniques in both undergraduate and postgraduate healthcare programmes, there is a consequent need to ensure the quality and rigor of these assessments with useful, defensible and legitimate statistical means such as psychometrics to detect and reduce or minimize such bias.

Errors can arise from the test, tester and testee. In OSCEs, errors are created in its production and interpretation and by processes impacting on the testing environment. These errors include ambiguous tasks that are too long or too short, invalid tasks and non-homogeneous tasks that are too hard or too easy, poor instructions, noisy exam venues, limited time for tasks, poor test design or variation in judges, patients or examination procedures, poor level of lighting and poor responses (Tavakol & Dennick 2011b:454).

Tavakol and Dennick (2011b:454) refers to the tester as "the person responsible for using and interpreting the assessment criteria". Errors from the tester include lack of understanding of assessment principles or item construction, lack of training in applying assessment criteria, lack of understanding of learning objectives, poor interpretation of assessment criteria, inconsistent application of assessment criteria, inconsistent scoring systems or mark schemes, sexist/racist bias, systematic typing errors, inter-rater variability and subjectivity in scoring. According to Tavakol and Dennick (2011b:454), a testee is the person being tested and is prone to the following errors: stress, illness and therapy, lack of and inconsistent teaching, poor learning environment, lack of appropriate resources, lack of practice opportunities, lack of sleep, poor emotional state, as well as poor communication and language skills.

Turner and Dankoski (2008:577) comment on some of the flaws of the conventional long and short case method of clinical evaluation such as subjectivity, non-uniformity, limited scope and number of patients per student in large exams which decreases the validity of the examination. Hence many medical institutions have adopted the OSCE though still retaining the conventional method. This new method has been largely applauded in well-resourced medical institutions globally. However, in many medical schools of limited resources, the OSCE experience and practice may differ. There have been several criticisms about the OSCE including its lack of depth and being difficult to implement in terms of cost, time, labour and human resources which might compromise the quality of its administration and scores, especially in resource constrained schools.

In the case of medical training in KIU-DAR-DAR, the programme is divided into four levels with a total of eleven semesters: basic sciences (one semester) in the first level, biomedical sciences (three semesters) in the second level, pathology (one semester) in the third level and clinical clerkships (six semesters) in the fourth level. The courses of one level are a pre-requisite for the next level in a spiral fashion. A semester lasts for between seventeen to twenty weeks. The clinical clerkships are divided into junior, junior-special, senior and senior-special in this order. The junior clerkship (two semesters) rotates through medicine, surgery and community medicine in one semester and Paediatrics, psychiatry, obstetrics and gynaecology in the next semester. This is followed by one semester of junior special clerkship in Dental Surgery, Ophthalmology, ENT, Anaesthesia, Forensic Medicine, Ethics in

Medicine and Radiology. The senior clerkship begins in the ninth semester with medicine and surgery followed by Paediatrics, Obstetrics and Gynaecology (OBGYN) in the tenth semester. In the eleventh semester, the students go for senior special clerkship in the specialized areas of Medicine and Surgery.

Students are assessed at the end of each semester and level. The end of semester examination comprise of 40% of the total assessment, whilst the promotional and exit examinations comprise of 60% of the total assessment. These examinations consist of written examinations, (multiple choice questions, short essays and long structured essays) and OSCE. KIU-Dar adopted the OSCE as a major method of clinical assessment in addition to the conventional (long and short case) about 48 months ago. The major emphasis of the faculty is on clinical training and assessment. Reports from various internship training centres and the community as well as complaints from the Uganda National Intern committee show that the quality of medical graduates from all universities is declining. In addition, the faculty has noticed some decline in the zeal of students towards active learning, quality training and assessment.

The use of OSCE has not been extensively researched in the region of Africa. Therefore, there are questions pertaining to the quality of the OSCE that the medical schools in this region implement. From the researcher's observation and having been involved in OSCE assessment for four years, the quality of the OSCE implemented in the low development countries have to a large extent not been measured objectively because the users are not conversant with the tools and techniques. Traditionally, KIU-Dar invites external examiners to evaluate the OSCEs. However, without standardized and objective psychometric tools, their judgments are often subjective, biased and inconsistent.

Based on the above-mentioned background to the problem, there is a need to re-evaluate our exit and promotion examinations. Incorporating globally more objective, cost-effective, defensible and standardized measures of assessment evaluation into our educational quality assurance system, which will be illustrated in this study, will reveal areas in our OSCE practices and training that need adjustment, thus improving our training and the quality of our graduates.

This study shed more light on the use of available psychometric methods in the analysis and interpretation of raw OSCE scores, thus making it a more readily available, user-friendly method of test evaluation which can be adopted by the faculty for subsequent use.

Feedback from this analysis and interpretation led to necessary corrective measures to improve OSCEs in KIU-Dar. Regular evaluation of our OSCEs over time will improve the quality of our examinations, training, graduates and community health care.

1.3 PROBLEM STATEMENT

The problems mentioned above about quality issues of the OSCEs in KIU-DAR-DAR require application of sound quality assurance measures to regularly monitor and evaluate this method of assessment for continuous improvement. One of these measures can be provided by adopting psychometric analyses into the university's quality assurance system for examinations. Also, Pell *et al*. (2010:802) identify other challenges facing resource constrained universities in this regard: many of the publications on this subject are either not accessible to the faculties in resource limited universities, contain only one or two methods at a time or have made understanding of this subject even more difficult. Moreover, the software to carry out these analyses is largely out of the reach of most educators in these regions. Thus, this study might assist in addressing the gap in or lack of knowledge, understanding and skill in quantitative psychometric analysis of the OSCE by illustrating simple step-by-step, comprehensive ways of measuring the most important and all-encompassing elements in the evaluation of the OSCE. At the end of this study, the illustrations are compiled into a guideline for the purpose of routine psychometric analysis of the OSCE in the institution.

Therefore, the problems that were addressed in this study include: the gap in or lack of knowledge and the complexity of the various available psychometric measures of analysis in order to answer the question of how to evaluate our OSCE objectively and comprehensively. The metrics of the scores obtained from a live OSCE in KIU-Dar, which is a typical resource limited medical school were provided in this study in response to the question of 'what is the quality of our current OSCEs in a typical resource limited centre?' In the interpretation, discussion of the findings and recommendations, the path to improvement was charted.

As Pell *et al.* stated 'Understanding and utilizing metrics effectively are therefore central to measuring quality and in directing resources to appropriate further research and development of the assessment' (Pell *et al.* 2010:802). There are very limited studies addressing psychometric analysis and as far as could be ascertained, none of these studies have been done in the regions of East and West Africa.

1.4 THE CONCEPTUAL FRAMEWORK

As mentioned earlier in Kamran *et al.* (2013b:e1464)'s report, psychometric analysis is an integral part of the quality assurance process of an assessment system that takes place following the OSCE. Other measures accompanying this analysis are: external examiners' reports, standardization, peer-review of items (stations), examiners' training and evaluation (cf. Figure 1.1). Some of these measures take place before the conduct of the OSCE. Of note is that the quality assurance of each examination is a continuous process, repeated with each examination cycle for the progressive improvement of the assessment exercises of an institution'. The figure shows the six components of a quality assurance system for an examination cycle.

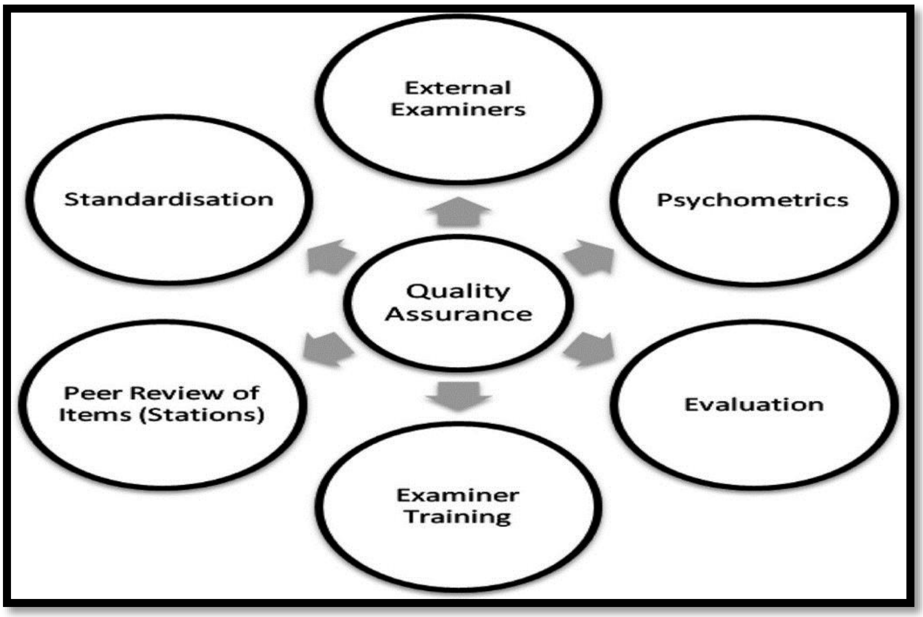

FIGURE 1.1: THE QUALITY ASSURANCE PROCESS [COMPILED BY PEL *et al.*, AMEE GUIDE NUMBER 49, 2010]

1.5 OVERALL GOAL OF THE STUDY

The overall goal of the study was to conduct a critical analysis of the OSCE using its post examination scores and therefore develop a guideline on post-examination psychometric analysis of the OSCE for subsequent use in institutions. This guideline may also be used to assess the quality of other objective examinations. Comprehensive psychometric methods which include the Classical Test Theory and the Item Response Theory were applied in a simplified manner to quantify, analyse and interpret the metrics of the OSCE scores in a resource limited medical school in East Africa, thus ensuring quality of assessment in the region

1.6 AIM OF THE STUDY

The aim of this study was to ensure quality in assessment of the OSCE scores using available psychometric methods in a resource limited institution. In the end, this study provides an illustration and then a guideline on the use of the relevant available traditional and one of the advanced psychometric methods to analyse and interpret the post-examination OSCE scores for quality assurance in a resource limited institution.

1.7 RESEARCH QUESTIONS

In order to address the problems stated above, the following research questions were addressed.

i. What should the psychometric methods that objectively and comprehensively evaluate the OSCEs with a view to ensure quality in an institution with limited resources look like?
ii. How should available psychometric methods of analysis of an OSCE be objectively and comprehensibly applied with a view to ensure quality in an institution with limited resources at KIU-Dar?
iii. How can psychometric methods be applied to analyse and interpret the raw scores of an OSCE at a medical school?
iv. What is the quality of the OSCE currently practiced in a typical resource limited medical school at KIU-Dar?
v. How can the findings of a psychometric analysis and interpretation be used to improve the OSCE at KIU-Dar over a period of time?

1.8 OBJECTIVES OF THE STUDY

Based on the above-mentioned research questions, the objectives of the study were to:

i. Describe the available psychometric methods for the OSCE through the following literature review.

The Classical Test Theory, which includes: (i) descriptive statistics (mean, mode, median and standard deviation), (ii) passing scores, (iii) number of failures, (iv) Cronbach's alpha 'total' and 'if the item (station) is deleted', (v) Item difficulty index, (vi) item discrimination index, (vii) statistical significance and (viii) standard error of the mean (SEM).

The Item Response Theory (IRT): The generalizability study.

ii. Develop a guideline for psychometric analysis and interpretation of the OSCE scores in a medical school.
iii. Illustrate the application of the available psychometric methods on the raw scores from a live summative OSCE in a resource limited medical school at KIU-Dar to obtain its psychometric properties and provide an appropriate interpretation of these properties.
iv. Describe the quality of the OSCE that is currently practiced in a typical resource limited medical school at KIU-Dar.
v. Describe how the findings of the analysis and its interpretation can be used to improve the OSCE at KIU-Dar in Dar es Salaam over a period of time.

1.9 DEMARCATION OF THE FIELD AND SCOPE OF THE STUDY

This study was conducted in the field of Health Professions Education (HPE). Psychometric analysis as a quality assurance measure for assessments including the OSCEs is a rapidly developing field that has become popular in medical education over the past few decades in the developed countries. However, there is a huge gap in the knowledge, comprehension and skills of psychometric analysis and the complexity of the various available psychometric measures of analysis is a problem in the various medical schools of the Sub-Saharan Africa. OSCE itself is relatively new in Africa, coupled with the fact that it is costly to implement standard OSCE and these have been a challenge to the quality of the OSCE implemented in many resource limited medical schools in Africa. The traditional over-dependence on external examiners in these medical schools, to vet examinations has been found to be subjective, biased and inconsistent in many cases. Therefore, psychometric analysis as one of the components of the quality assurance system of the examination cycle provides an objective, stable, quantifiable measure of the quality of the OSCE.

In the current field of HPE, psychometric analysis is an important measure that should be adopted for regular use to improve assessments in educational programmes and therefore the quality of medical graduates and health care provision to the community. A thorough literature review of psychometric analysis of the OSCE was done and a practical illustration of its application was carried out using a live OSCE in a resource limited medical school in Tanzania.

The results of the study can be applied to medical programmes of other medical schools in East Africa and the rest of Africa with similar challenges in terms of financial, human resource and technical support. The results can also assist with other forms of assessments and in other faculties. The study was conducted between November 2012 and June 2016, with the empirical research phase from June to July 2015.

1.10 SIGNIFICANCE, CONTRIBUTION AND VALUE OF THE STUDY

Currently, there is no formal guideline to objectively measure and monitor the quality of the OSCE and other forms of assessments in the researcher's university and other learning institutions in Africa. Hence, the significance of this study lies in its ability to bridge the gap of lack of awareness and skills in psychometric analysis on the part of examiners, because the contents of this study will be made accessible to the trainers in resource limited schools through publications in conferences, seminars, workshops and journals. Moreover, the simple but comprehensive illustrations of psychometric analysis in this study, will improve the understanding of trainers. Subsequently, the OSCE practice in these medical schools will improve steadily, resulting in quality medical graduates and patient care. The value of this research will ultimately reside in the development of a guideline for psychometric analysis for the improvement of the OSCEs and the integration of psychometric analysis into the current quality assurance system of the examination process in KIU-DAR-DAR as a required component and not only as an optional activity. The School of Health Sciences at KIU-DAR-DAR is open to educational projects such as that offered in this study to assist in the improvement of its OSCEs and other examinations, for the enhancement of education and training for undergraduate students.

1.11 HYPOTHESIS

The null hypothesis tested in this study was that the psychometric properties of the OSCEs implemented in a typical medical school in Sub-Saharan Africa are within the good ranges (H_o=good ranges) despite its limited resources. The alternative hypothesis is that the psychometric properties of the OSCEs implemented in a typical medical school with limited resources in Sub-Saharan Africa are below the good ranges (H_a<good ranges). Therefore the statistical analysis that will be applied on the scores that will be gathered in this study will be one tailed because the alternative hypothesis stated is directional (Araoye 2003:37-38).

1.12 RESEARCH DESIGN OF THE STUDY AND METHODS OF INVESTIGATION

This study adopted a descriptive cross-sectional research design, which implies that the scores for analysis were compiled in one OSCE examination setting (Araoye 2003:55). Descriptive research is a process of collecting data in order to answer questions concerning the current status of the subjects in the study and in this case, the researcher described the current psychometric properties of the OSCE practiced in a typical medical school with limited resources in East Africa (Borrego, Elliot & Amelink 2009:54). This study on psychometric analysis and interpretation of the OSCE scores adopted a quantitative approach. The quantitative elements in this study were supposed to be the primary numerical and global scores from an implemented OSCE.

The researcher observed the OSCE proceedings with a checklist, without intervening in the process. The numerical and global scores for the students were to be recorded on checklists by the examiners and provided on mark sheets by the clinical departments. The global score is the examiner's personal professional judgment of the student's overall performance in his station irrespective of the numerical scores he has given based on the sum of marks on the checklist. Global scores were to be recorded on a 5-point Likert scale as: excellent, pass, borderline, poor, fail (Pell *et al.* 2010:807). Data on the opinions of the standardized patients involved in the OSCE about each candidate's performance were to be collected using semi-structured questionnaires. Statistical and text analysis were carried out using Microsoft Excel and SPSS computer packages (Lennon-Dearing & Barnes 2013: 8 of 23).

1.13 DESCRIPTION OF THE METHOD

The researcher obtained the available psychometric methods for the OSCE by literature review. The live OSCE set-up (station design and station contents) and proceedings were described by the researcher together with two assistants and the detailed OSCE scores per examiner per candidate per station were obtained on excel-sheets after the examinations from the clinical heads of departments. The psychometric analysis was carried out on the scores obtained from each station (examiner) and total (candidate) as well as overall students' performance. The post-OSCE scores were subjected to the following statistical analysis in order to determine the reliability of the scores:

1.13.1 *The psychometric analysis*

The following steps were taken in this study to carry out psychometric analysis on the OSCE station scores:

1.13.1.1 *Providing general information about the subject*

General information about the examiners, students and subject stations were first documented to assist in subsequent analysis.

1.13.1.2 *Frequency distribution*

A descriptive analysis was done to summarize and present the raw scores of the OSCE. The frequency distribution showed any missing scores and the pattern of scoring. Histogram charts were created to show the pattern of distribution of the scores (Kuzma & Bohnenblust 2001:31). Z-scores were generated from the normalised raw scores. The following five components were used to describe the shape of the distribution of the station scores.

1.13.1.3 *Measures of central tendencies*

Measures of central tendencies were computed to identify the central score for each station.

1.13.1.4 *Measures of variability*

Measures of variability were computed to determine the variation of the scores within and between the stations.

1.13.1.5 *Station analysis*

The quality properties of the stations were described.

1.13.1.6 *Reliability checks*

Reliability estimates were used to describe the consistency, correlation and importance of the stations.

1.13.1.7 *Identifying Hawks and Doves*

The following three steps were followed to identify Hawks and Doves:

- In Step 1, potential extreme examiners were identified by comparing an individual rater's mean score to the mean of all raters for that station.

- Step 2 involves comparison between the distribution of ratings from the extreme examiners identified in Step 1 and the distribution for all examiners to determine whether the examiner demonstrated adequate variability in their candidate ratings for a given station.
- Step 3 determines whether the candidate cohort seen by the examiners in question demonstrated adequate variability.

1.13.1.8 *Guideline for psychometric analysis*

The guideline for psychometric analysis was developed for subsequent use.

All of the above tests were under the Classical Test theory except for the generalizability test which falls under the Item Response Theory (IRT). The generalizability tests were applied to identify the sources of errors and measure the effects, the various factors of the OSCE have on the students' performances. The details of the psychometric analysis carried out in this study are discussed extensively in Chapters 2 and 3.

1.14 IMPLEMENTATION OF THE FINDINGS

The research findings will be submitted to the management committee of the School of Health Sciences, KIU-DAR-DAR, and of the Faculty of Health Sciences, University of the Free State, with recommendations that the findings of this study could be used to:

- Train the university academic staffs in psychometric analysis of examination scores as one of the quality assurance measures.
- Integrate psychometric analysis into the current quality assurance system of the examination cycle.
- Improve other forms of examinations using psychometric analysis in the university.
- Present in seminars, workshops and conferences.
- Publish in journals.
- Improve medical training.
- Improve quality of medical graduates.
- Improve patient safety.

The results of this research can play a proactive role in the need to upgrade educational techniques to health care education. The regular use of psychometric analysis in the examination process in the various programmes in the university can contribute significantly to the improvement of the quality of our graduates and patient care in the future.

The research findings will be submitted to academic journals, for publication as articles. This study should make a meaningful contribution to the use of psychometric analysis to improve the OSCEs especially in resource limited medical schools. The findings of the research will be brought to the attention of other medical schools in East Africa and the rest of Africa and can be adapted or used as such by these schools as a guide for the implementation of psychometric analysis as a quality assurance measure in their examination process.

1.15 ARRANGEMENT OF THE REPORT

The report of the research findings and the final outcome are arranged as follows:

Chapter 1 introduces and gives a background to the study.

Chapter 2, **Psychometric analysis as a quality assurance system in the OSCEs**, in turn explores the psychometric analysis as a quality assurance measure for the OSCEs and a literature review of the psychometric analysis, its role in the education and training of health care professionals. This chapter also deals with the educational and practical implementation of psychometric analysis in the examination cycle in the medical programme.

Chapter 3, **Research design and methodology**, gives the methodological approaches employed to study the research problem.

In Chapter 4, **Results of the psychometric analysis of the OSCEs,** a report of the results of the observations of the researcher and the psychometric analysis of the OSCE scores is given.

In Chapter 5, **Interpretation and discussion: Psychometric analysis of the OSCEs,** we give an interpretation and discussion of the study findings.

Finally, Chapter 6, **Conclusion, recommendations and limitations of the study**, gives the conclusions, recommendations and limitations of the study.

CHAPTER 2

PSYCHOMETRIC ANALYSIS AS A QUALITY ASSURANCE SYSTEM IN THE OSCEs

2.1 INTRODUCTION

This chapter provides an overview of assessments, a description of psychometric analysis and the documented psychometric properties of the OSCE as seen in some medical schools worldwide. The literature on assessments, OSCE, quality assurance system of the examination cycle and psychometric analysis was reviewed in order to develop the theoretical basis for the empirical part of this study.

2.2 ASSESSMENTS

According to Aranda and Yates (2009:2), assessment refers to all the methods used to determine the extent of an individual's achievement of learning outcomes. Reinert (2013:25) classifies assessment methods into cognitive and behavioural according to Miller's pyramid which was introduced into medical education literature in 1990. Crossley *et al.* (2002a:800) and Cruess, Cruess and Steinert (2016:180) describe the four categories of assessment methods in Millers pyramid, based on what is required of the trainee. Furthermore, the authors noted that the measures of cognition fall into the lower level of the Millers pyramid and are collections of 'Knows' and 'Knows how' assessment methods, whereas the 'Shows how' and 'Does' categories of assessment methods were rated high in the pyramid and were assessments of behaviour. The higher levels assessments, also referred to as performance-based assessments, build upon the lower levels.

Reinert (2013:25) declares that it is impossible to assess the clinical competencies of the trainee without simultaneously assessing his knowledge. Although the highly rated assessment methods are considered to more closely reflect professional reality, they are not superior to the lower levels of assessment in Miller's pyramid (Cruess *et al.* 2016:180). The authors affirmed that the superiority of assessments lies on how best suited the assessment is to the purpose of the test. Reinert (2013:25) also acknowledges the complexity of the tasks involved in the assessment of the behaviour of the medical student in the clinical phase of their training.

Furthermore, Barman (2005:479) points out that a comprehensive assessment method for the medical student in clinical training should be able to test all the core competencies of the trainee effectively. However, Miller (1990:S63) has earlier stated that no single assessment method can provide all the data required for judgement of all these core competencies equally and therefore had highly recommended the use of a combination of multiple methods of assessments which can overcome many of the limitations of an individual assessment method. The best approach is to utilise several appropriate methods of assessment longitudinally to test all the core competencies in the medical student in stages.

Regarding trained competencies, the Accreditation Council of Graduate Medical Education (ACGME), America (Barman 2005:480) has documented the following six core competencies to be acquired by the trainee:

- Patient care: Which is the ability of the trainee to provide compassionate, appropriate and effective patient care for the treatment of health problems and promotion of health.
- Medical knowledge: The ability of the trainee to demonstrate knowledge of the established and evolving biomedical, clinical, epidemiological and social behavioural sciences as well as the application of this knowledge to patient care.
- Practice based learning and improvement: This is the ability of the trainee to investigate and evaluate ones care of patients, to appraise and assimilate scientific evidence and to continuously improve patient care based on constant self-evaluation and life-long learning.
- Interpersonal and communication skills: This competence results in the effective exchange of information and collaboration with patients, their families and health professionals.
- Professionalism: The trainee should demonstrate a commitment to carrying out professional responsibilities and an adherence to ethical principles.
- Systems-based-practice: The trainee should demonstrate awareness of and responsiveness to the larger context and system of health care and also have the ability to call effectively on other resources in the system to provide optimal health care.

Barman (2005:479) recognises that performance-based assessments, unlike cognitive achievement-based methods are more psychometrically limited because, it is more difficult to design and administer the performance-based assessments to the level of standard that is considered fair and effective especially in resource limited institutions. Therefore, it is advisable that performance based assessment methods are better utilised to screen for minimal level of competence or to identify outstanding trainee performance. In the author's opinion, the key quality features of assessment that need close attention are validity and reliability. However, since reliability is a necessary component of validity, investigating the reliability of a high stake assessment using psychometric analysis is of utmost importance for ensuring quality in assessment. One of the performance based assessments that is currently commonly used in medical schools globally is the OSCE. This form of assessment is elaborated below.

2.3 OSCE

OSCE was first described in 1970 by Harden and has been refined over the years to the level that it is now highly appreciated by the majority of stakeholders including trainers and trainees. For example, in Mccrorie and Boursicot (2009:224)'s report of the graduate examinations used in the medical schools in the UK, the authors documented that 25 out of the 32 medical schools use OSCE. Similarly, Rozycki (2003:1) reports that since 2005, the US National Board for Medical Examination requires all graduating medical school students in the United States to pass an OSCE as part of the US Medical Licensing Examination. OSCE is also being used in medical schools in Australia, New Zealand, parts of Europe, Africa and Asia.

Despite its popularity, OSCE has a major psychometric limitation in the area of content specificity, non-consensus in scoring, poor inter-ratter reliability, time consuming and resource intensiveness (Mccrorie & Boursicot 2009:224). Hence, there is a pressing need to regularly measure the quality of the OSCE, objectively, using psychometric methods, especially in resource constraint schools such as KIU-Dar.

2.4 QUALITY ASSURANCE SYSTEM OF THE EXAMINATION CYCLE

Pell *et al*. (2010:804) describe the six components of the Quality Assurance system of the examination cycle, which can be implemented before or after the examination. These components include: external examiners' reports, standardization, peer-review of items (stations), examiners' training, psychometrics and evaluation. Standardization of pass mark can be done before and after the examination. Psychometric analysis takes place after the examination. The quality assurance of each examination is a continuous process, and therefore should be repeated with each examination cycle for the progressive improvement of the assessment exercises of an institution. Mccrorie and Boursicot (2009:226) raises some quality issues in their study regarding the absence of a national licensing examination for the medical profession, whereby, every medical school in the UK decides on its own systems of assessment, including its graduating examinations. In other words, graduation from any UK Medical school and the conferment of a university degree in medicine usually leads automatically to the granting of a licence to practise by the medical professional regulatory bodies. The assumption, as stated by the authors, is that the examinations at all the UK medical schools guarantee the same levels of minimum clinical and professional competence required for licensure and practice. However, Mccrorie and Boursicot (2009:226) found that there was an extensive variation of assessment processes to such an extent that it was difficult to compare the equivalence of standards of graduates from the different medical schools. Their findings include variable number and expertise of external examiners at the different schools suggesting that there may indeed be inconsistencies in the standard of assessment in the different medical schools. Previously, McManus, Elder, De Champlain, Dacre, Mollon and Chis (2008:2) published the significantly varied levels of performances of graduates from different UK medical schools in all parts of the MRCP (UK) examinations. Moreover and Wass (2005:791), in her article, campaigned energetically for a national licensing process to commence in the UK. It appears that this quality issue is not only restricted to resource-poor institutions but is also experienced in medical schools of developed countries and perhaps at different levels of magnitude.

A similar situation exists in Vietnam and Asia whose first experience of the OSCE was in 2006 at Hue Medical school with the assistance of the Path finder organization and since then the use of OSCE has spread to other medical schools in the region (Fan, Tran, Kosik, Mandell, Hsu & Chen 2012:103). In the same vein, the authors commented on the absence of a national licensing examination in Vietnam, which raises serious quality issues. Hence, to correct these quality flaws in these two and other regions with similar arrangement, the authors recommend the introduction of a national licensing examination, which all graduates have to pass in order to be licensed to practice or better still to employ a more objective method of measuring the quality of our assessments such as using psychometric methods.

In the US, Europe and UK, where psychometric methods have been employed, Barman (2005:480) recorded a reliability coefficient of the OSCE ranging between 0.41 to 0.88 in some of the medical schools. This occurred, despite stringent control over the assessment process. However, the search for published work on psychometric properties of the OSCEs conducted in medical schools in Africa has been largely unfruitful.

2.5 PSYCHOMETRIC ANALYSIS

In Tavakol and Dennick (2011b:447)'s view, the psychometric analysis of OSCE stations has been less reported in the literature in comparison to knowledge based tests that measure learning. Medical educators need to measure how much material has been learnt by their trainees by evaluating the results of particular achievement tests. The authors further believe that analysing examination questions and scores by means of psychometric methods may provide useful information to improve the OSCE. The authors emphasised that efforts should be directed to minimising all errors influencing a test in order to produce an observed score which approaches a learner's true score as reliably and validly as possible. As earlier stated, reliability is the most important measurement characteristic of an assessment because its analysis reveals the consistency, usefulness and practical value of the test. The procedures that were utilized in the measurement of the quality of the OSCE in this study include the following.

2.5.1 Observations on the day of OSCEs

Pell *et al.* (2010:803) stress that OSCE assessment should be systematically observed. It is possible to anticipate and correct, in advance of the OSCE, many of the contributing factors to error variance. According to Pell *et al.* (2010:803), the following points should be observed with a checklist during the OSCE:

- Ensure similarity in design across the stations.
- Ensure that the stations follow the requirements of the checklist design.
- Ensure that the set-up of existing parallel OSCE circuits match.
- Ensure that stations carry the same provision of equipment (or permit flexibility if students are taught different approaches with different equipment).
- Check the time assessors arrive and that they attend the pre-assessment briefing. Late coming assessors might miss the briefing and therefore fail to adhere adequately to the prescribed methodology.
- Look out for unauthorized prompting by assessors.
- Look out for inappropriate behaviour or excessive interaction by assessors.
- Look out for excessively proactive simulated patients whose questions may act as prompts to the students.
- Look out for biased real patients (e.g. gender or race bias). Many real patients may not be able to undergo training to the same level undertaken with simulators on how

to interact with the candidates. Hence this limitation might negatively influence the candidates' performances.

- Look out for assessors (or assistants) that do not return equipment to the start or neutral position as candidates change over.

2.5.2 Station length and number

The quality of the OSCE is influenced by the number of stations in the OSCE test and its duration. The size of many exams is based on tradition or length of time available rather than appropriate sample size which will ensure that the score obtained by the student reflects their global knowledge. Steven (2011:85) claims that OSCEs usually require at least 4 hours of testing for them to be reliable. However, the challenges associated with such lengthy examinations include costs, acceptability to students, difficulties with organization and examinee tiredness which may ultimately affect the psychometric properties of the assessment. According to Steven (2011:85), the appropriate sample size for OSCE can be calculated using the formula below:

$$N= \frac{Z^2\,(SD)^2}{E^2}$$

Where **N is the number of stations**.
Z is the confidence level indicating how much the sample size is influenced by chance (1.64 for 90% confidence, 1.96 for 95% and 2.57 for 99%). Usually 95% confidence level is taken.
SD is an estimation of the standard deviation in the population of items.
E^2 is the type I error of the sample size usually set at 0.05 or 5%.

(Mohsen & Reg 2011:448)

The station number was computed with Microsoft Excel by the statistician in this study. The accepted number of stations ranges from 8 to 20, but on the average 15 for good reliability.

2.5.3 Standard setting

The pass mark for each station and the overall pass mark for the assessment should be decided by standard setting, before or after the OSCE. In addition, and independent of any scoring, assessment procedures should include a global judgment by the examiners of pass, fail or borderline (Norcini 2003:464). The methods for deciding the pass mark could either be relative (norm-referenced) or absolute. The absolute methods could either be based on the test item (criterion) or the performance of the candidate (borderline methods). The author recommends the absolute standards, because they are defensible, evidence-based, acceptable, and the most commonly used.

From Tavakol and Dennick (2011b:450)'s experience, a large number of institutions currently favour the use of the borderline regression method (BLR) for determining pass marks, as only the BLR method uses both the global grade and checklist score to investigate the relationship between the two and also the level of discrimination between weaker and stronger students (Hejri, Jalili, Muijtjens & Van Der Vleuten 2013:891). The BLR method needs some level of expertise for computation and uses five global ratings (e.g. fail, borderline, pass, credit, distinction). The authors' concern about the BLR method is that it is very sensitive to outliers and they categorised the outliers in three main groups:

- Students who perform very badly and obtain a near zero checklist score.
- Students who achieve a creditable checklist score but who fail to impress the assessor overall.
- The assessor who gives the wrong overall grade.

However, in the situation where global scores and expertise is lacking, the simple borderline method can be used to determine pass marks.

In this study, the global scores were not provided by the examiners, hence, the statistician determined the pass mark for each station, student total and for the overall assessment using borderline method with Microsoft Excel. In the borderline method, each station scores were arranged in ascending order and then divided into three equal groups. The mean of the middle group is computed and presented as the pass mark for that station. This borderline pass mark is closely related to the median of the scores in the station (Pell *et al.* 2010:807).

The student *t*-test was also used to identify significant discrepancies between the generated station pass mark and the fixed university pass mark.

2.5.4 *Generating station level quality metrics*

According to Tavakol and Dennick (2011b:452), once the raw scores have been obtained from an OSCE, the stations are sorted according to the blueprint and their scores entered systematically into the statistical programme for psychometric analysis in the following order.

2.5.4.1 *Frequency distribution*

The frequency distribution of scores

Inspection of the distribution graphically with a histogram can reveal how far the scores deviate from a 'normal' distribution and how skewed they are.

Skewness: Describes the horizontal shape of the scores distribution.
Kurtosis: Describes the vertical shape of the scores distribution.
Outliers: Unusual extreme scores which may be large or small and suggests errors in the data.

Z-scores are obtained after normalising raw examination scores, so that they can be compared with other scores from other examinations in a standard way. The Z-scores can be calculated from the means and standard deviation by the equation:

$$\boxed{z= \frac{X-\ddot{X}}{S}}$$

A Z-score is equal to the difference between a raw score (X) and the mean score of students (Ẍ) in a particular test divided by the standard deviation (s). All Z-score transformed distributions have a mean of 0 and a standard deviation of 1. An individual's z-score shows how far above or below the mean their score is in units of standard deviation. The researcher will illustrate this by using an example from Tavakol and Dennick (2011b:452): assuming the mean of scores in a particular test is 50 with a standard deviation 15. If a student scores 65, his/her Z-score is +1. This means that the student is +1 standard deviation above the mean of the distribution. Standard tables of Z-scores are available for comparing the position of students to each other.

Within the normal distribution the position of scores is as follows: 68% of scores lie within ±1 standard deviation of the mean. Ninety-five percent of scores lie within ±2 standard deviations of the mean. Finally, 99.75% of the scores lie within ±3 standard deviations of the mean. Therefore, referring to the above example as further explained by the authors, approximately 16% of other students obtained higher scores than the student in the example. Thus, relying on a raw score can provide a wrong impression of the student, as well as a distorted view of the exam. Z-scores allow teachers to compare students' scores on different tests with different total scores (Tavakol & Dennick 2011b:452). These statistics can be computed with SPSS and Microsoft Excel.

2.5.4.2 *Measures of central tendencies*

There are commonly three measures of central tendencies and these are explained below.

Mean: This is the average of all the scores.
Mode: This is the most frequently occurring score in the distribution.
Median: This is the midpoint of the distribution, where 50% of the scores fall on either side.

Comparing the sum of the scores in each station could reveal some aspect of the symmetry of the scores distribution. Tavakol and Dennick (2012:e163; 2011b:451) reports that the differences between the mean, mode and median give a more objective indication of how much the distribution deviates from the normal. The skewness of a test scores' distribution indicate the overall ease or difficulty of the test. When the mode is off to one side the distribution is said to be skewed. If the mode is to the left with a long tail to the right the distribution has positive or right skewness. This shows that few students' test scores fall at the high end of the distribution, which means the test was too difficult. In order to modify the discrimination at the lower end of the distribution, more tasks with a lower level of difficulty should be used. If the long tail is to the left, the distribution has negative or left skewness and this shows that few students' test scores fall at the lower end of the distribution; this implies that the test was too easy. In this regard, the authors advise that, in order to modify the discrimination at the higher end of the distribution, harder tasks should be used.

2.5.4.3 *Measures of variability*

The measures of variability either measure the differences in the scores within the station or between the stations. These measures are elaborated below.

Range

Range refers to the difference between the maximum and minimum scores in a data. Because of its limitations, range is not commonly used to measure variability in a data.

Standard deviation

Standard deviation (SD) is the square root of the variance and gives an indication of the spread and variability of the data. This is a useful and relative measure of variation or dispersion. On the one hand, Tavakol and Dennick (2011b:452) interpret a low SD as an indication that the examination was either too easy or too difficult and that there is little variability in the examination. On the other hand, a high variability with a mean at the centre of the distribution indicates an examination that is very useful.

Coefficient of variation in percentage

The coefficient of variation is an absolute and very reliable measurement of variation within the station. The coefficient is calculated by dividing the sample (station) standard deviation by its mean and expressed in percentage. Therefore, unlike the other measures of variability which are relative, it can be used to compare variability in several stations (Kobayashi, Sakuratani, Abe, Yamazaki, Nishikawa, Yamada, Hirose, Kamata & Hayashi 2011:64).

Standard Error of the Mean (SEM)

The SEM provides an estimate of the amount of error inherent in an individual's test score and determines the discrepancies between an individual's observed score on the test and his/her true score. There is an inverse relationship between the test reliability estimate and the SEM, meaning that the larger the test reliability estimate, the lower the SEM (Tavakol & Dennick 2011b:456). The method of computing the standard error of the mean in SPSS is shown in Appendix E.

95% Confidence Interval of the Mean

This was generated by SPSS in the study as a measure of variability of the stations. The wider the 95% confidence interval of the mean, the more unreliable are the scores in the station (Kuzma & Bohnenblust 2001:117).

Homogeneity Test

The Homogeneity test assesses the uniformity of the tasks across all the stations. In this study, the Levene test in SPSS was used to test for the homogeneity of the station tasks (Pell *et al.* 2010:807).

Analysis of variance (ANOVA)

The ANOVA measures between and within-group variations. According to Pell *et al.* (2010:807), all the variation in scores in an ideal OSCE process will be due to differences in student performance or intrinsic abilities, and not due to differences in other factors between or within the groups which are external to the student such as the environment (e.g. local variations in layout or equipment), location (e.g. hospital-based sites having different local policies for management of clinical conditions) or differences of assessor attitude (i.e. hawks and doves). ANOVA can be used to measure the effect of each of these factors on the students' performances where the assessment arrangement is complex. ANOVA is useful in this set-up whether the groups or assessors or sites and other factors are randomly or non-randomly selected. Basically, ANOVA determines whether there is a significant difference between or within samples and if there is, to identify which of the sample(s) is different from the others.

In the context of this study, the one-way ANOVA was used to investigate the differences between the groups of students' test scores in multiple OSCE stations with different tasks in each subject as well as overall. This is a very powerful metric as it gives a very good indication of the uniformity of the assessment process between and within the OSCE stations. Ideally the between station variances should be under 30%, and values over 40% should give cause for concern, indicating potential problems at the station level due to factors external to the student.

Generalizability Theory Analysis

According to Tavakol and Dennick (2012:e172), reliability estimates, such as Cronbach's alpha, cannot identify neither can it discriminate between the potential sources of measurement error or facets associated with a test. However, the Generalizability Theory or 'G-theory' which is an extension of CTT, under the Item Response Theory (IRT), attempts to recognize, estimate and isolate these facets allowing test constructors to gain a clearer picture of sources of measurement error for interpreting the true score. The G-theory was developed by Lee J. Cronbach and colleagues in 1972 (Tavakol & Dennick 2012:e172). Unfortunately, it appears that most investigators are not familiar with the generalizability theory because of the absence of analytic facilities for this purpose in popular statistical software packages. Each facet of measurement error has a value associated with it, called its variance component, calculated via an analysis of variance (ANOVA) procedure, described below. The variance components give the percentage of variance specific to each of the factors influencing the checklist scores of the students' performances. These variance components are next used to calculate a G-coefficient which is equivalent to the reliability of the test and also enables one to generalise students' average score over all facets.
The SPSS can calculate the variance components directly from the test data. The procedure used varies according to the number of facets in the test. Tavakol and Dennick (2012:e172)

stated that there are single facet and multiple facet generalizability designs. The authors described the two designs in this manner: for a single facet design, only a single source of measurement error in a test is examined, but in reality others may exist.

For example, in an OSCE examination, the focus might be on the influence of examiners as sources of error. In G-theory, this is called a one-facet 'student (s) crossed-with-examiner (e)' design: (s × e).

The authors further explained that in a multi-facet design, several sources of measurement error are examined, such as, the number of stations, the number of SPs and the number of items on the OSCE checklist. The scores for each of the facets under investigation are inserted into the G-study programme for analysis (Tavakol & Dennick 2012:e172).

After entering the examination data into the SPSS, the data are restructured from a multivariate design to a univariate design, whereby the primary outcome variable of interest (i.e. ratings) appears in only one column. Univariate formats will result in data sets with multiple records (rows) per object of measurement (e.g. persons), whereas multivariate formats will typically have only one record per object of measurement (Putka & McCloy 2004:2 of 24). After restructuring, then analysis can commence.

In SPSS, the analysis of variance components is carried out as described by Tavakol and Dennick (2012:e172): 'To this end, from the data menu at the top of the screen in SPSS, one clicks on 'restructure' and follows the appropriate instructions. Then to obtain the variance components, the following steps are carried out: From the menus choose 'Analyse', 'General Linear Model', respectively. Then click on 'variance components'. Click on 'Score' and then click on the arrow to move 'Score' into the box marked 'dependent variable'. Click on student and examiner to move them into 'random factors'. After 'variance estimates' appears, click OK and the contribution of each source of variance to the result will be presented'.

In the terms of G-theory, according to Mushquash and O'Connor (2006:542), the variance of the students is not considered a facet of measurement error as this variation is expected within the student cohort and it is called the 'object of measurement'. Another component of the G-theory is the residual variance, which according to Nunnally and Bernstein (1994:22 of 774) is the amount of variance not attributed to any specific cause but is related to the interaction between the different facets and the object of measurement of the test. The G-coefficient (p^2) is defined as the ratio of the student variance component (denoted 'Vs') to the sum of the student variance component and the residual variance (denoted 'Ve') divided by the number of examiners (k) and written as follows:

$p^2 = Vs/Vs + (Ve/k)$ for a single facet design. In a multi facet design, $p^2 = Vs/Vs +(Vi/k+Ve/k+Vst/k+Vsp/k)$

Where,
p^2 = G-coefficient
Vs = Student variance component
Ve = Residual variance
K = Number of examiners
Vi = Examiner.student interaction variance
Vsp = Standardized patient variance
Vst = Station variance

(Tavakol & Dennick 2012:e173)

The G-coefficient (ρ^2) is also a reliability coefficient with values ranging from 0 to 1.0. According to Tavakol and Dennick (2012:e173), the G-coefficient in the single facet design described above is equal to Cronbach's alpha coefficient. The interpretation of the value of the G-coefficient is that it represents the reliability of the test taking into account the multiple sources of error calculated from their variance components. The higher the value of the G-coefficient, the more we can rely on or generalize the students' scores and the less influence the study facets have been on the students' performance.

2.5.4.4 *Item (Station) Analysis*

Station or Item analysis describes the quality properties of the stations. Tavakol and Dennick (2011b:452) documented that item analysis of test results uses quantitative methods to help make judgments about which stations need to be adopted, which stations need to be revised and which stations should be discarded. Item analysis determines the ease or difficulty of individual stations as well as the relationship between individual stations and the global station score. It includes the item difficulty index and the item discrimination index. In the example that these authors give, if students with high scores overall do well in a particular station, the tasks in that station would be considered to be good. Equally, if a student with low scores overall performs poorly in a particular station, the tasks in that station would be considered to be good.

Item (Station) Difficulty Index (IDI)

A station, where all the students who were examined, either passed or failed, is not a good one and needs revision. The tasks in that station are either too easy or too difficult for a student and therefore contribute little information regarding the student's ability (Tavakol

& Dennick 2011b:452). According to the authors, the item-difficulty index or item facility refers to the percentage of the total number of students who passed the station and is calculated as follows:

Pi = Ri/N

Where,
R = The total number of students who passed the station
N = The total number of students who were examined at that station (passes + failures + no response).
i = The station number.
P = The fraction of the students that passed at the station.

In the authors' illustration, if 40 out of 100 students passed station 'one', the item difficulty index is simply calculated as follows:

P1 = 40/100 = 0.40

The value of an item-difficulty index ranges from 0 (if no one passed the station) to 1 (if everyone passed the station). Hence, the larger the P value, the easier the station. The good range of Pi is located between 0.3 and 0.8. This statistics was computed with Microsoft Excel.

Station-Discrimination index *(d)*

The station-discrimination index, as described by Tavakol and Dennick (2011b:453), is a value of how well a station is able to differentiate between students who are high performing and those who are not, or between 'strong' and 'weak' students. The range of *d* is -1.00 to +1.00, but good *d* is between 0.3-0.5. The method of *d* recommended by the authors is explained thus: the examiner divides students into two groups ('high' and 'low') according to the score sheet of each student. On the basis of this classification, 27% of the students are categorized as a strong group and 27% as a weak group. Hence, 46% of the middle-scoring students are excluded from the calculation of the item-discrimination index. Next, the number of students (in both groups) who answer a particular question correctly is calculated. The following formula is used to calculate a *d*-value:

d = (U - L)/n

Where U equals the number of the students who passed the station in the upper group, L = the number of the students who passed the station in the lower group and n is 27% of the total number of students (Tavakol & Dennick 2011b:453). To illustrate further, the authors used this example: 'in a physiology test of a total of 112 medical students. The top and bottom 27% of the test scores, with a total of 28 students in each group were isolated. It was observed that 18 students in the 'strong' (top) group passed the station and 10 students in the 'weak' (bottom) group passed the station. Therefore, the *d*-value is equal to:

0.28= [(18–10)/28]'

A negative *d*-value on a given station indicates that the 'strong' students failed the station and the 'weak' students passed the station. The authors advised that such stations should either be revised or discarded. This statistics can be computed with Microsoft Excel.

Statistical significance

Statistical significance is another method of station discrimination index. Tavakol and Dennick (2011b:453) describe the procedure in this manner: The students are divided into two groups, those who passed the station, called 'group R', and those who failed the station, called 'group W'. The mean of the total score of 'group R' and 'group W' is calculated. The mean score of group R (ẌR) could be below or above the mean score of group W (ẌW). Consequently, the null hypothesis that should be considered is that 'ẌR is equal to ẌW', weighed against the alternative hypothesis that 'ẌR is greater than ẌW'. The null hypothesis means that there is no difference between the mean scores of the students who passed the station and those who failed it. To test the null hypothesis, a t-test can be used to assess whether the means of two groups (ẌR and ẌW) are statistically different from each other or not. If the p-value is less than 0.05, we will reject the null hypothesis and accept the alternative hypothesis. This means that the station has divided students into two separate strong/weak groups (Tavakol & Dennick 2011b:453). This statistics was computed with Microsoft Excel in this study.

Number of failures

According to Pell *et al.* (2010:807), the global rating from the expert judgement of trained assessors against the expected performance of the minimally competent student is used to determine the failure rate. An unusually high number of failures does not always indicate that a station is too difficult. Failure rates may be used to review the impact of a change in teaching in a particular topic. A value more than 10% indicates where a review of content and methods of teaching, can improve course design.

2.5.4.5 *Reliability checks*

The reliability tests describe the consistency, correlation and importance of the stations. Reliability is concerned with the reproducibility, stability and internal consistency of an assessment (Tavakol & Dennick 2011b:454). In the authors' opinion, the internal consistency of a test is a measure of how well the individual stations are functioning together to measure the same underlying constructs and how accurately and precisely it is able to measure the construct of interest. Reliability is the key estimate showing the amount of measurement error in a test. Test reliability is a function of the difference between the observed test score of the student and his/her 'true' score. The observed score is the score that a student obtains from an actual test. The true score is the score that a student obtains from a (hypothetical) test when it accurately measures his or her underlying ability. If there is a significant difference between an observed test score and a true score, the reliability of the test is low, and vice versa.

The authors further explained that reliability is concerned with the error inherent in psychometric measurements. Reliability is the correlation of the test with itself, squaring this correlation, multiplying it by 100 and subtracting from 100 gives the percentage error in the test. Using the authors' example, if an examination has a reliability of 0.80, there is 36% error variance (random error) in the scores. As the estimate of reliability increases, the fraction of a test score that is attributable to error will decrease. Conversely, if the amount of error increases, reliability estimates will decrease (Tavakol & Dennick 2012:e162).

According to Tavakol and Dennick (2011b:454), the factors that cause errors in measurements are classified as external and/or internal and may arise from three sources namely: the test, the testee (student) and the tester. The external factors depend on the test situations and administrations, such as the room temperature, guessing answers, emotional problems, physical discomfort, lack of sleep, scorers and scoring systems. The internal factors depend on the quality and quantity of the test, such as poor or limited item sampling and the way in which the item is constructed. Reliability of an OSCE can be determined by two broad theories: the classical test theory and the item response theory.

The classical test theory seeks to identify and minimize the sources of error in a test. In many cases errors can be identified and controlled before an assessment is undertaken but it is practically impossible to anticipate or estimate every possible error. As a result, the determination of the true reliability coefficient of a test is not practicable (Tavakol & Dennick 2011b:454). The reliability estimate of a test is obtained from the data acquired after the test has been administered using the techniques for estimating reliability described below:

Coefficient alpha

Alpha was developed by Lee Cronbach in 1951 (Tavakol & Dennick 2011a:53). Alpha is widely used for estimating the internal consistency reliability or station homogeneity. Alpha reveals the amount and effect of measurement errors on an observed test score of a student cohort rather than on an individual candidate (Tavakol & Dennick 2011b:455). Alpha is a correlation of a test with itself. On the one hand, the students who receive the same score on a homogenous OSCE station have a similar knowledge in the area tested. On the other hand, those who receive the same score on a heterogeneous OSCE may have different knowledge in the areas tested (Tavakol & Dennick 2011b:456). Thus, the test scores that come from a heterogeneous station are more ambiguous than a homogenous test. The authors revealed that the homogeneity of a station is also an indicator of construct validity as it ensures that all the questions on the test, measure the same construct or trait. Alpha can be considered as an estimate of the interrelatedness of a set of test's items, whereby in a good assessment the better students should do relatively well across the board (i.e. on the checklist scores at each station). Alpha can be applied to both dichotomous and continuous variables. Alpha ranges in value from 0 to 1 and should be above 0.70, but not much more than 0.90. If the alpha is above 0.90, it may suggest that some stations are redundant (testing the same task but in a different guise) and the test length should be shortened (Tavakol & Dennick 2011b:455).

The authors further explain that in a unidimensional test, the (overall) value for alpha can be reported; but in an heterogeneous test, alpha 'if item deleted' should be calculated for the group of stations or station of each case. Hence, for a good unidimensional test, the alpha is from 0.7 and above. But for a heterogeneous test where each station metrics are standardised, a higher alpha would be expected, from 0.754 but not more than 0.90. In the situation where alpha is >0.90, it is an indication of unnecessary duplication of content across stations and points more to redundancy than to homogeneity. Another indication of redundancy in the stations is when the alpha 'if item deleted' scores are higher than the overall alpha score. Alpha tends to increase with the number of stations in the assessment regardless of whether the test is homogeneous or not. A low overall alpha indicates poorly designed stations or can sometimes be attributed to large differences in station mean scores. Generally, the alpha 'if item deleted' scores should all be lower than the overall alpha score if the item/ station has performed well. Where this is not the case, this may be caused by any of the following reasons (Tavakol & Dennick 2011b:455).

- The station is measuring a different construct to the rest of the set of items.
- The station is poorly designed.
- There are teaching issues – either the case being tested has not been well taught, or has been taught to a different standard across different groups of candidates.
- The assessors are not assessing with a common standard (Tavakol & Dennick 2011b:455). The method of computing Cronbach alpha is shown in Appendix E-IV.

Pearson's (Interclass) correlation, *r*

Pearson's test is the correlation between the mean station scores and the mean total OSCE scores and measures the direction and strength of the linear relationship between variables with numerical values. The higher the *r* value, the better the station is at discriminating. The *r* values range from -1.0 to +1.0. A station with a negative *r* should be revised or discarded. In Tavakol and Dennick (2011b:453)'s opinion, the *r* test increases the homogeneity of an OSCE exam. The r^2 estimates the importance of the station and its contribution to the variation in the students' scores. r^2 is obtained by squaring *r*. The guideline for the Interpretation of correlation coefficients was documented by Kuzma and Bohnenblust (2001:125) as follows: '0.75-1.00: strong; 0.50-0.74: moderate to high; 0.25-0.49: low to moderate; 0.00-0.24: weak'. The method of computing the Pearson's (Interclass) correlation in SPSS is shown in Appendix E-IV.

Coefficient of determination (R^2) and Inter-grade discriminative tests

R^2 coefficient is the proportional change in the dependent variable (checklist score) due to change in the independent variable (global grade). R^2 shows how good one term is at predicting another and determines the degree of linear correlation between the checklist score and the overall global rating at each station, with the expectation that higher overall global ratings should generally correspond with higher checklist scores (Pell *et al.* 2010:805). The overall global rating is the examiner's professional judgement of the student's performance regardless of the checklist scores. The Inter-grade discrimination index indicates how the average increase in checklist mark corresponds to an increase of one grade on the global rating scale. The inter-grade discrimination index gives the slope of the regression line and is therefore closely related to R^2. According to the authors, the discrimination index should be of the order of a tenth of the maximum available checklist mark. The square root of R^2 is the simple Pearson correlation coefficient (Hejri *et al.* 2013:891).

A good correlation ($R^2>0.5$) will indicate a reasonable relationship between checklist scores and global grades. However, Pell *et al.* (2010:805) warns against examiners that directly translate the checklist score into the global grades, thereby artificially inflating R^2. Hence, the authors recommend the plotting of a scatter graph or SPSS curve estimation of checklist scores against global ratings as routine good practice, regardless of station metrics. Unsatisfactory relationship between checklist scores and global ratings will cause some degree of non-linearity. From the authors' example, when some students have acquired high scores from the item checklist, but with fail global grade, the slope of the regression line will increase. The reverse occurs when the checklist scores are poor but the global grades are reasonable. The correlation curve can either be linear (good correlation), quadratic or cubic. The authors argued that in mathematical context, a cubic curve will always produce a better fit, but parsimony dictates that the difference between the two fits has to be statistically significant for a higher order model to be preferred. The authors pointed out that the cause of the cubic expression must be investigated and could result from an underlying relationship between the checklist scores and the global grades or as a result of outliers, resulting from inappropriate checklist design or unacceptable assessor behaviour in marking. The final judgment on the relationship between the checklist scores and the global grades is determined by the distribution of scores seen on the scatter graph, the correlation curve and the stations' metrics. Where stations' metrics are generally of good quality, a departure from strict linearity is not a cause for concern. The existence of low R^2 values at certain stations and/or a widespread of scores for a given grade should prompt a review of the item checklist and station design (Pell *et al.* 2010:805). The method of computing R^2 in SPSS is shown in Appendix E.

2.5.4.6 *Standardised patient ratings*

In Pell *et al.* (2010:808)'s experience, simulated/standardized patients (SPs) are often required to rate candidates, and this typically follows an intensive training programme.

Assessment in simulated, as opposed to real-life, clinical settings can be limited by the Hawthorne effect. The Hawthorne effect is the tendency of research subjects to behave atypically as a result of their awareness of being studied, as opposed to behaviours occurring as a result of the actual treatment that has occurred. This SP rating therefore assesses what the learner is capable of rather than what they actually do in the real-life setting.

SPs can be asked a question such as 'Would you like to consult again with this doctor?' With a range of responses (strongly agree, agree, neither agree nor disagree, disagree or strongly disagree) using the five-point Likert scale, the two latter responses being regarded as adverse. A higher than normal proportion of candidates (e.g.>10%) receiving adverse SP ratings may indicate problems. The SP ratings should not be used alone but be reviewed alongside other station metrics, and the impact on SP ratings monitored in response to training or other interventions. If this is coupled with a higher than normal failure rate, it could be the result of inadequate teaching of the topic. The overall reliability of the assessment may be increased by adding the SP rating to the checklist score. Typically, the SP rating should contribute 10–20% of the total station score (Homer & Pell 2009 in Pell *et al.* 2010:808).

2.5.4.7 *Identifying Hawks and Doves*

Researchers have estimated that 12% of variance is due to differences between examiners in leniency-stringency in the British clinical examination (McManus*,* Thompson & Mollon 2006:7). In the USA, reports on the clinical skills exam used to assess international medical graduates (IMGs) have also shown similar variance estimates. For example, Boulet (2003:27) reports an associated value of 10.9% and Roberts (Roberts, Rothnie, Zoanetti *et al*. 2010:690) reported a value of 8.9%.

In general, variability due to examiners is considered to be highly undesirable with potential negative impacts on the validity of scoring outcomes. The phenomenon of examiners being on two ends of the scoring spectrum i.e., some of the examiners scoring very leniently and others scoring very harshly, has been recognized since the beginning of the 20th century (Harasym, Woloschuck & Cunning 2008:618), and it is often referred to as the Dove/Hawk phenomenon. Dove, used in English as a term of endearment, describes lenient examiners. In addition, Hawk, used in English as a portrayal for any form of predator, describes stringent examiners. But, Thorndike (1920:25; McManus, Thompson & Mollon 2006:40 of 57) describes the two main categories of examiner errors, namely: correlational and distributional errors. Correlational errors occurs where raters demonstrate a tendency to evaluate or rate an examinee holistically and similarly across the rating scales, without discriminating between different dimensions of behaviour or performance, which is called the halo effect. The distributional errors exist where raters fail to make adequate use of the full range of a rating.

McManus *et al.* (2006:8 of 57) defines two common distributional errors which are range restriction and leniency/severity errors. Leniency/severity errors occur when a rater systematically rates candidates too kindly or too harshly. Systematic measurement error due to raters, according to Raymond and Viswesvaran (1993:254) occurs when the mean of a rater summed over all the candidates differs significantly from the mean of all other raters. Bartman, Smee and Roy (2013:29) designed a simple three steps procedure of classifying examiners as Hawks and Doves to pick out distributional errors, as follows:

- Step 1 - Identify potential extreme examiners by comparing an individual rater's mean score to the mean of all raters for that station. In Step 1 all examiners whose *average* score for a station was more than three standard deviations above (potential dove) or below (potential hawk) the average score for all remaining examiners on that same station is a potential extreme examiners. The three standard deviations identify the very extreme ratings and are usually a very small percentage of the examiner population. The ANOVA was used to identify potential extreme examiners in this study.
- In Step 2 the distribution of ratings from extreme examiners identified in Step 1 were compared to the distribution for all examiners to determine whether the examiner demonstrated adequate variability in their candidate ratings for a given station. The analyses in Step 2 have a double purpose.
  - The first purpose is to eliminate the possibility that the stations' unique scoring key causes the extreme scoring. For example, stations with a small range of possible scores may engender more extreme scoring than stations with wide range of possible scores. The coefficient of variation was used to determine the variability in each station for this study.
  - The second purpose is to evaluate if the extreme rater is able to discriminate among the candidates, at least to a certain degree. The examiner might have failed everybody, or almost everybody, yet he/she may still have assigned higher scores to better candidates and lower scores to worse candidates. In this case, the extreme examiner's rating is still providing valuable information regarding differences in candidate performance and such an examiner is thus eliminated from further

scrutiny with regard to their scoring practices. The item discrimination index, *d* was used for this purpose in this study.

- In Step 3, the final step in identifying extreme raters, the cohort criterion, is to determine whether the candidate cohort seen by the examiners in question demonstrated adequate variability. For example, if a large majority of candidates rated by a potential dove consistently performed higher than average on the other stations, then we would not classify the examiner as a dove. Similarly, if most of the examinees that a potential hawk rated were poor candidates then we might no longer classify the examiner as a hawk. Here, we use data from all stations to determine candidate ability overall and to isolate extreme judgments compared to other examiners (Bartman, Smee & Roy 2013:29). In this study, the Pearson's correlation, *r* was used for this final step.

2.6 POST EXAMINATION REMEDIAL ACTION

In Pell *et al*. (2010:809)'s experience, it is highly unlikely that any institution would re-examine candidates even with poor OSCE metrics. Therefore, action needs to be taken to ensure that all assessment decisions are defensible, equitable towards students and rigorous from an institutional perspective. The authors emphasised teamwork between clinical academic staff work and psychometricians in deciding how to remediate the effects of poorly designed examinations. He proposed a number of pragmatic, post examination remediation methods, when faced with unsatisfactory metrics, as follows:

Adjustment of total scores to a common mean and ensure the failing students are not confined to a single site. This is usually the easiest method. Adjustment at the station level is seldom carried out because at the time of totalling the stations' scores, any adverse effects tend to cancel each other out.

Removal of a station is also a rare event and according to the author, the criteria for this are usually multiple adverse metrics, the result of which would disadvantage students to such an extent that the assessment decisions are indefensible against appeal (Pell *et al*. 2010:809).

Other post-examination redesigns of the OSCE, according to the authors include:

- Chunking of a number of simple criteria into fewer criteria of higher level on the assessor checklist.

- The inclusion of intermediate grade descriptors on the assessor checklists.
- Ensuring that checklist criteria have three instead of two anchors where appropriate, thereby allowing greater discrimination by assessors.
- Ensuring a greater degree of uniformity between the physical arrangements of the different circuits.
- Upgrading of assessor training methods.
- Updating ('refreshing') assessors who were trained some time ago.
- The provision of more detailed support material for assessors.
- Improved assessor briefings prior to the assessment.
- Improved SP briefings prior to the assessment.
- Dummy runs before the formal assessment for both assessors and SPs (this is only really practicable where student numbers are relatively small, e.g. resits.

(Pell *et al*. 2010:809)

In closing, the authors commented on the need for the application of a family or battery of metrics on the OSCE so that a true picture of quality can be obtained and the deficient areas identified. All the above improvements would be unlikely to have been apparent from using a single reliability metric, such as Cronbach's alpha or the G-Coefficient. Adopting this approach will be rewarded with a steady improvement in the delivery and standard of clinical assessment (Pell *et al*. 2010:810).

2.7 CHAPTER CONCLUSION

In this chapter, an overview was given of what the literature states on the following aspects: Assessments: Cognitive and Behavioural methods of assessments. OSCE: Quality Assurance System of the Examination Cycle. Psychometric analysis, as part of the quality assurance system and post examination remedies.

The literature review was conducted with the view to form the theoretical basis and to support the empirical part of this study.

In Chapter 3, the methods used to conduct this study will be discussed.

CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

3.1 INTRODUCTION

In this chapter, the research design and the methodology applied in this study are described. The theoretical orientation and explanation of the design and selected methods of data collection will be discussed. A descriptive cross sectional research design was used to guide the research. In addition, the data collecting methods and tools, namely, a literature review, observation OSCE checklist, scores checklists and semi-structured questionnaires are described. Furthermore, the study site, target population, sample, sampling technique and sample size will be described. Also included in this chapter are the statistical techniques used for data analysis, anticipated limitations, assumptions, validity and reliability as well as the ethical considerations.

3.2 THEORETICAL PERSPECTIVES ON RESEARCH DESIGN AND METHODOLOGY

The theoretical perspectives that follow will describe the research design and the research methods.

3.2.1 Research design

A research design is a plan or procedures of enquiry that specifies how the research is going to be executed in such a way that it answers the research question (Creswell 2014:31). The research design for this study was non-experimental, quantitative, descriptive and cross-sectional. A non-experimental research is a systematic empirical investigation in which the scientist does not have direct control of independent variables because their manifestation has already occurred or because they are inherently not manipulable (Thompson & Panacek 2007:18). On the other hand, experimental research seeks to obtain answers by manipulating a condition or introducing some change into a situation (Cook, Beckman & Bordage 2007:737). This study is non-experimental, as there was no attempt to manipulate any variable.

Non-experimental research designs can either be qualitative or quantitative. Quantitative studies are studies that use mathematical analyses that can reveal statistically significant differences. Qualitative studies refer to studies that focus on theory or logic and use open-ended questionnaires and they do not normally produce numeric data (McDaniel & Gates 2001:98; Creswell 2014:42). This study is quantitative, as a lot of mathematical analysis was carried out.

Quantitative studies are of three general categories, which are causal-comparative, descriptive and experimental. Causal-comparative studies reveal a systematic cause and effect relationship between independent and dependent variables, whereas, a descriptive study systematically document current measurable events. During experimental research, treatment of an intervention introduced into a study group is investigated and then the outcome of the treatment is measured. (Jones & Bartlett 2007:19; Williams 2007:66). This study was aimed at describing the quality of the OSCEs currently implemented in an example of a research limited institution and the psychometric measures utilised. Accordingly, the study is descriptive in nature. In addition, this study is cross-sectional by design (Institute for work and Health 2015:1 of 2) as data was collected at a single point in time during the OSCEs.

3.2.2 Data collection methods

A data collection method refers to the general strategy followed in gathering and analysing the data necessary for answering the questions at hand (Araoye 2003:55). The research questions and objectives were discussed in Chapter 1, while the research results will be presented in Chapter 4. In this section the researcher explored the methods and tools of collecting data, the sample of the target population and the data analysis strategy employed. The data collection strategies in this study were literature review, observation and the questionnaire methods. The data collection tools utilised were the OSCE observation checklist and the assessor's checklist. A semi-structured questionnaire was designed for collecting data from standardised patients during the OSCE (cf. Appendix EIII). However, in this July 2015 final OSCEs, the departments did not recruit nor train any standardised patient, hence the questionnaires were neither served nor used for this study.

Data collection tools are specific instruments that are appropriate for or accompany and facilitate the chosen data collecting method to accomplish its tasks in a proper way. For example, the observation method is facilitated by checklists, the questionnaire method is facilitated by questionnaires (Mangal & Mangal 2013:313). In this study, three data collecting tools were utilised: the researcher's checklist for observing OSCE design and proceedings, the examiner's checklist and the standardised patient's questionnaire for observing student's behaviour during the OSCE. In designing these data collection tools for this study, the different levels of measurement scales were put into consideration. Garger (2010:1 of 3) explains that there are basically four types of measurement options. These measurement options are nominal, ordinal, interval and ratio scales. Nominal scales are the lowest level of measurement, where variables are classified into unordered qualitative categories. Ordinal scales classify data into ordered qualitative categories, for example social class (I, II, III, etc.) where the values have a distinct order, but their categories are qualitative in that there is no natural (numerical) distance between their positive values. In an Interval scale, information about the ordering of magnitude of the measurement and about the distance between the values is conveyed. Likert scale questions such as the agreement scales used in the current study classify as interval scales. While these are strictly speaking ordinal in nature, they are often considered as interval scales by researchers to enable the calculation of means and parametric significance testing. Ratio scales have equal distance between the numbers, as with interval scales, yet it also has an absolute zero. The components of the data collection methods employed in this study are discussed as follows.

3.2.2.1 *Literature review of psychometric methods for the OSCE*

According to Cronin, Ryan and Coughlan (2008:37), 'a literature review is an objective, thorough summary and critical analysis of the relevant available research and non-research literature on the topic being studied'. The aim of literature review is contextualising the problem against current related theory and research, as well as to ensure that the researcher is sufficiently knowledgeable about the topic to be able to investigate the topic in an informed manner. In Bowen's (2005:210) opinion, the researcher should read extensively, so as to be equipped with a wide range of current literature and work towards a particular topic. This helps the researcher to determine gaps and areas where further research is needed. For this study, electronic databases containing primary research studies such as Medline, Pubmed and Biomed Central were the major resources consulted. Search for articles was also conducted in Research gate. Other resources containing secondary or synopsis of primary studies utilised include textbooks, official reports and thesis (Krupski, Dahm, Fesperman & Schardt 2008:1264). The collection of references consulted is listed in the bibliography section of this study. The literature review of the psychometric methods used in this study was discussed in Chapter 2. The researcher was able to readily access literatures on psychometric methods used on the OSCEs of medical schools in developed countries. However, it was very difficult to get published articles on the OSCE and its evaluation in medical schools located in West, East and Central Africa.

3.2.2.2 *Observation method*

The observation method was utilised to collect information about the OSCE design and proceedings, so as not to influence nor interrupt the OSCE proceedings in this study. Observation method is a way of collecting data through observing (i.e. watching what people do) and can be structured (controlled) or unstructured (natural). Other sub-classifications include overt/disclosed (where the participants know they are being studied) or covert/undisclosed observation (where the researcher hides his real identity from the research subjects, acting as a genuine member of the group). Unstructured observation, on the other hand, is conducted in an open and free manner with no pre-determined variables or objectives (McLeod 2015:1-7). According to literature, the observation technique used to collect information on OSCE design and proceeding in this study was structured (with an OSCE checklist) and overt. The observation was overt, because the researcher explained the research aim to the group, so the participants knew they were being observed. Controlled observations are usually non-participatory as it was in this study, because the researcher avoided any direct contact with the group, keeping a distance (McLeod 2015:1-7).

Observation method is relatively cheap and requires few resources. The researcher has direct access to the research information and is able to collect primary or original data, which is more reliable. In observation, knowledge is acquired through the use of the sense organs, only relevant things are taken into account and thus it is essentially selective, flexible and objective (Araoye 2003:56). Observation in this study was carried out using the eyes, ears and mind to perceive the ongoing activities during the OSCE. Closer attention was paid to the relevant information required or variables which are being measured and then recorded in the checklists as the case is, in structured observation. The checklist is a pre-defined schedule prepared for guiding and recording observations (Baker 2006:172).

In structured observation, the checklist enables easy replication by other researchers, thereby making it easy to test the reliability of the study. The data obtained is easier and quicker to analyse as it is quantitative compared to naturalistic observations, which is qualitative. Moreover, controlled observations are quick to conduct which means that a large sample can be obtained within a short time and results in findings that are representative and can be generalized to a large population (Kawulich 2012:6 of 20).

Subjective bias of both observer and respondent is eliminated since information pertains to what is happening and is gathered with a pre-defined checklist. Furthermore, unlike the interview method, observations are independent of respondents' willingness to answer questions. This method is useful especially when the respondents are not capable of answering verbal question as in an examination such as the case in this study. The flaws with this method include the Hawthorne effect, which may reduce the validity of the study. The Hawthorne effect exists, when participants act differently because they know they are being watched. The information provided may be limited and sometimes unforeseen factors may interfere with the observation. In some cases respondents may not be willing to be observed (Araoye 2003:56; Sedgwick & Greenwood 2015:1 of 2).

A brief description of the steps in conducting a structured observation is as follows:

Determine the focus of observation: Because of typical time and resource constraints, the observation exercise has to be selective, looking at a few activities that are central to the research questions; hence the focus of observation must be clear. In this study, the researcher observed the OSCE design and proceeding in four clinical departments, whilst the examiners observed the behaviour of the students as they carry out the tasks in the stations.

Design observation checklist: After determining the focus of observation and before starting a structured observation, a checklist of items to be observed as well as the recording structure is designed to facilitate easy analysis and increase reliability (cf. Appendix EI & EII). The observers in this study were the researcher assisted by two volunteers to increase validity. The volunteers (non-experts) were trained in the use of the research instruments (checklist and questionnaire) and how to conduct observational method of data collection before piloting. The research procedure and instruments were revised in the light of the pilot test results and minor changes were made in the researcher's observation checklist before the main study was carried out

The next step is to determine the site of observation. Sites are usually selected based on expert's advice. In this study, selection of the venue for the OSCE was not the responsibility of the researcher. Kampala International University, Dar es Salaam (KIU-Dar) does not have its own teaching hospital yet, but depends on Tanzania government hospitals for its clinical training and assessments. The Faculty of Medicine (FoM), KIU-Dar selects the hospital for the OSCE, based on existing memorandum of understanding between the two organisations.

Timing is critical, especially when events are to be observed as they occur. Wrong timing can distort findings. The timing for the OSCE was decided by the faculty and the university according to the university academic calendar for examinations. Observations were carried out during the OSCEs in this study.

Next, field observation is conducted. The researcher should make effort to establish rapport with the population to be studied. The presence of outside observers may generate some anxiety among those being observed. Establishing rapport is often achieved by creating informal, friendly conversations to reduce anxiety levels. Informing the participants of the purpose of the observation is not to report on individuals performance, but to find out what kind of problems in general are being encountered, also helps to create a conducive atmosphere for the exercise. In this study, the researcher had obtained permissions and consent to carry out the research from relevant authorities and had briefed the examiners and student-participants about the research before the OSCE commenced (Sommer & Sommer 1986 in Northern Arizona University 2001:10 of 14; Mookherjee, Monash, Wentworth & Sharpe 2015:4 of 7).

In the field, sufficient time should be allowed for observation. Brief visit can be deceptive, partly because people tend to behave differently in the presence of observers. It is not uncommon for example, for health workers to become more caring or for extension workers to be more persuasive when being watched. However, if observers stay for relatively longer periods, people become less self-conscious and gradually start behaving naturally. It is essential to stay at least two or three days on a site to gather valid, reliable data (Halder, Molyneaux, Luby & Ram 2013:13 of 17). In this study, the researcher and assistants were with the participants throughout the four OSCE sessions, which took place daily for four days. A team approach was utilised during observation. The three observers (researcher and two assistants) observed the OSCE design and proceedings in each of the four departments together. Recording of observations during the OSCE, was carried out as inconspicuously as possible, so as not to distract the examiners and students. As a team, a more comprehensive, high quality data, devoid of individual bias was obtained.

Checklists for observation during OSCE

Checklists are paper-based tools for recording observations in research. Video-recording can also be used to record observations. These checklists help to standardise the observation process and ensure that all important items are covered. The checklists also facilitate better aggregation of data gathered from various sites or by various investigators

(Araoye 2003:57). When preparing a structured observation checklist as the case is, in this study, it is necessary to consider the following:

- Identify in advance the possible response categories for each item, so that the observer can answer with a simple yes or no, or by checking the appropriate answer. Closed response categories help minimize observer variation, and therefore improve the quality of data.
- Limit the number of items in a form. Forms should normally not exceed 40-50 items. If necessary, it is better to use two or more forms than a single large one that runs several pages.
- Provide adequate space to record additional observations for which response categories were not determined.
- Use of computer software designed to create forms can be very helpful, as it facilitates a neat, clear checklist that can be easily completed.

(Verdaasdonk, Stassen, Widhiasmara & Dankelman 2009:715)

The content of the checklists used in this study were obtained from literature review of the OSCE and psychometric methods. The checklists' items were clear, straightforward and require close-ended responses, hence, it was easy to train non-experts to use it. Two types of checklists were used in this study for observation: one for the researcher and the other for the examiners. The researcher's observation checklist was prepared by the investigator, while that of the examiners was sourced from the Ministry of Health by the departments. These two types of checklists are described below.

Researcher's checklist for observation during OSCE

In this study, the researcher's checklist shown in Appendix EI, consists of items to be observed during the OSCE and the observer's responses. The responses are either in 5-point Likert scale, circle option/s or yes/no format (cf. Appendix EI). The data collected by this checklist is quantitative in nature with interval scale of measurement except for items i and ii, which are on the nominal scale of measurement. The content of this checklist was used to observe the OSCE design and proceedings in the four clinical departments.

Examiner's checklist for observation during OSCE

The examiner's checklist used in this study belonged to the Ministry of Health of Tanzania. This checklist is used for all levels of the medicine programme. Each of the clinical department (OBGYN, Paediatrics, Internal Medicine & Surgery) has its own specific checklist. The checklist has 11 sections. This checklist, however, does not cover global grades. The global grade is the examiner's personal professional judgment of the student's overall performance in the station, irrespective of the numerical scores the student has acquired based on the sum of marks on the checklist. Global grades were recorded (handwritten by the examiners) on a 5-point Likert scale as 'excellent, pass, borderline, poor and fail (Pell *et al.* 2010:807) in the checklist. The data collected by this checklist is quantitative discreet in nature with interval scale of measurement. The global grades are quantitative discreet on ordinal scale of measurement, however, they were not recorded in this study, because the examiners were not familiar with global scoring. The structure and marks allocated for each section of the examiner's checklist (total mark is 100) is shown in Appendix EII.

3.3 RESEARCH CONTEXT

The researcher works in and carried out this study in Tanzania, one of the countries in East Africa (cf. Figure 3.1). East Africa is one of the most challenged regions in Africa due to political and economic instability and disease. KIU-Dar es Salaam is one of the very few private-owned and the youngest university with a medical school in East Africa region and therefore, a good example of a resource limited medical school. The researcher is a staff of KIU, but the study findings were not affected by any conflict of interest.

3.3.1 Study site

The process of selecting KIU-Dar es Salaam, School of Health Sciences and its July 2015 final OSCE as the study site in this research has been explained in section 3.4. KIU-Dar is a multi-campus institution, where students can transfer their academic pursuit from one campus to another under the umbrella of student exchange program. Kampala International University began in 2001 in Kampala, Central Uganda, its first medical school started in 2004 in Ishaka, western part of Uganda. The KIU-Dar es Salaam (KIU-Dar), School of Health Sciences began in 2011 and the FoM (study site) emerged from the School in 2015, Figure 3.3. KIU-Dar is located in the Gongolamboto, the coastal region outskirt of Dar es Salaam region, which is the capital of Tanzania, East Africa (cf. Figure 3.1 & 3.2). The strength of the school for now lies in its biomedical sciences departments, while the clinical departments are in their formative stages of development. The MBCHB programme the university implements in the FoM, KIU-Dar has been described in section 1.2.

The current population of medical students in the faculty is 471, spread across all the levels of the MBCHB programme. Recently, the FoM, KIU-Dar was rated the best medical school in Tanzania by the East Africa Medical Council. The teaching hospital for the university is currently under construction (cf. Figure 3.4), but the university has memorandum of understanding with several Tanzania government hospitals for the purposes of training and assessment. One of such hospitals was selected by the FoM, through the university to conduct the final OSCE. The hospital was Amana regional referral hospital, and the OSCEs took place in the patient wards of OBGYN, Paediatrics, Internal Medicine and Surgery departments. The Amana Hospital is 30 minutes' drive away from the KIU-Dar campus, and is located in the centre of Dar es Salaam.

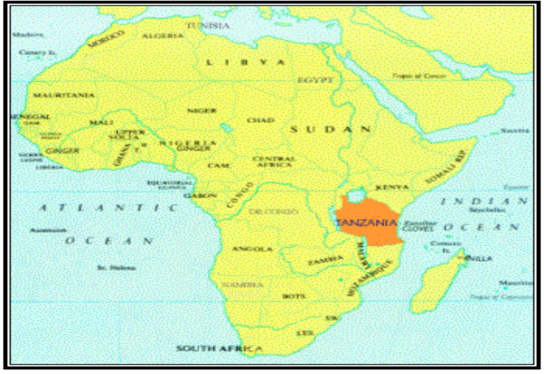

FIGURE 3.1: TANZANIA IN EAST AFRICA

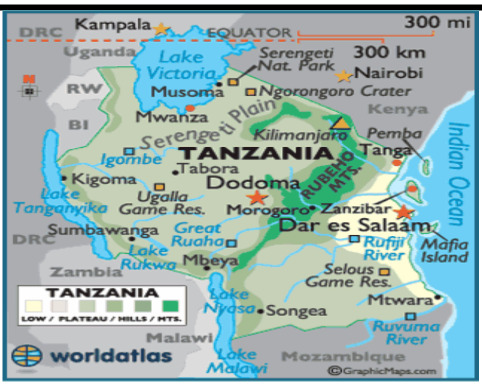

FIGURE 3.2: DAR ES SALAAM

KIU
EXPLORING THE HEIGHTS
Kampala International University
Dar es Salaam College

Enroll now for
PHD, Masters, Undergraduate, Diploma and Certificate programs

Available in
-Distance & Regular
-E-Learning (MBA)

Contacts:
+255685647070
+255715081028
E Mail: admin@kiu.ac.tz
Website: www.kiu.ac.tz
(Click this ad to visit our website)

FIGURE 3.3: FOM, KIU-DAR

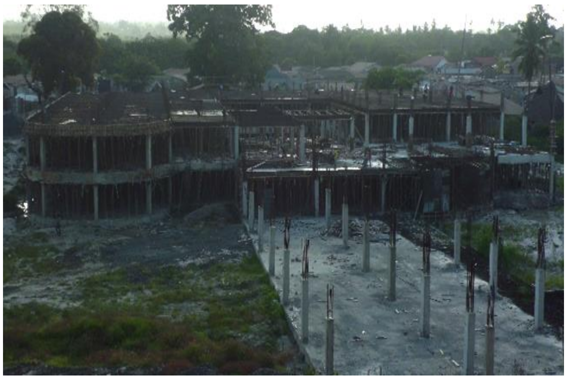

FIGURE 3.4: KIU-DAR, TEACHING HOSPITAL

3.3.2 *Study population*

The most advanced or high stake OSCE in the university was selected for this study. The most high stake examination in the school is the final examination because most of the university resources (human, time, finance) is invested in this examination as compared to others. Hence, determining the psychometric properties of the final examination will allow a fair generalization of the findings to other examinations in the same school. KIU-Dar es Salaam, for now graduates students only once a year every November. Hence, the final examinations are usually scheduled for July according to the university academic calendar. The empirical research of this study falls in 2015 according to the researcher's time schedule. Therefore, the July 2015 OSCE was selected for this study. The July 2015 OSCE was scheduled to take place over four days, one day for each of the four clinical departments (OBGYN, Paediatrics, Internal Medicine & Surgery).

A population is an entire group about which some information is required to be ascertained (Banerjee & Chaudhury 2010:60), which is the whole academic community and process in KIU-Dar es Salaam. Salkind (2000:86) defines the target population as a group of potential participants to whom the researcher wants to generalise the results of the study (Banerjee & Chaudhury 2010:60). The target population for this study was the entire medical students in the FoM, KIU-Dar, their examiners, their OSCE scores and standardised patients recruited for OSCE. An academic rule in FOM, KIU-Dar, is that a student is not allowed to sit for the OSCE of any subject, without first passing the theory of that subject. Other eligibility criteria for registering and sitting for examinations in KIU-Dar include 100% financial (tuition fees) clearance and 75% class attendance. Hence, the study population in this research was defined as, 'all the current and active clinical medicine students in their final year of FOM, who registered, sat for and passed their exit OBGYN, Paediatrics, Internal Medicine and Surgery theory examinations and were therefore eligible to sit for the July 2015 final OSCE of that subject'. The invited examiners that participated and the OSCE scores generated in the July 2015 OSCE were also considered as part of the study population for this research (cf. section 3.3.3). For this research, the potential study population was obtained from the Dean's register, which contained the list of all the current and active final year clinical medicine students, who were potential candidates for the exit OSCE in July 2015. Each student in the Dean's register was allocated a serial number for the purpose of identification throughout the OSCE in the four clinical departments. The students were not required to fill any questionnaire or checklist. The list of the confirmed study population (candidates who had passed the theory) was obtained on the day of each OSCE (cf. Table 4.2). The participating examiners were sourced by the heads of the clinical departments. The participating examiners were required by the departments to fill the examiner's checklist and the heads of departments consented to release the scores to the researcher after compilation of all the examiners' checklist scores and marking of the written stations was completed. In the July 2015 OSCE, there was no standardised patient because the departments did not have sufficient time to recruit and train the necessary standardised patients.

The excel copies of the OSCE scores were obtained from the heads of the four clinical departments, after the invited examiners have scored the students' performances on checklists and the written OSCEs have been marked by the medical officers in the various departments. The examiners submitted only their checklist scores. Despite the pre-OSCE briefing, the examiners did not record their global grades/scores because they were not familiar with it and global grades were not captured in the Tanzania Ministry of Health examiners' checklist. Psychometric analysis of these final examinations' results is a reflection of the quality of the graduates who sat for these examinations.

3.3.3 Sample, sampling techniques and sample size

A sample is any part of the fully defined study population (Banerjee & Chaudhury 2010:62). Trochim (2008:1of 1) defines sampling as the process of selecting units (e.g. people) from the study population so that by analysing the sample, a fair generalization of the results can be applied to the target population from which they were chosen. Umar and Folorunsho (2012:65) describe the two types of sampling designs, which are: probability and non–probability sampling methods. The non-probability sampling methods include the accidental/convenience/haphazard sampling, quota sampling and judgement sampling.

The sampling in which personal judgment and not randomness determines which units of the population are selected is called non-probability sampling. The accidental / convenience / haphazard sampling is accidental and convenient and therefore cheap and easy to use, but may lead to unreliable results. In this study, the entire study population was included because of its small size. Hence, from the above definition of sampling, no sampling was necessary in this study because the entire study population was recruited for the research (cf. Section 4.3, Table 4.2).

The minimum size of the data expected from this study population was:

- All the final year clinical medicine students in the faculty Dean's register book, expected to sit for the July 2015 OSCE = 30.
- Expected number of OSCE stations per department = 20.
- Minimum number of expected examiners per department = 2.
- Minimum number of expected standardised patient per department = 1.
- Number of departments = 4.

The amount of data expected from each department was:

- OSCE scores=200 data.
- Global scores= 40 data.
- Standardised patient= 20 data.
- Total=260.

Therefore, for the four clinical departments, the minimum number of data expected was 1024. In the actual study, there were a total of 82 OSCE stations, 27 students, 20 examiners, no standardised patient, no global scores. Hence, the total number of data collected across all the four departments from the study population was 2460.

3.3.4 *The pilot study*

A pilot study was done to ensure appropriate application of the psychometric analysis tools on the OSCE scores. The data collecting tools that were tested for validity, reliability, clarity and simplicity included the questionnaires for the standardised patients, the assessor's and the researcher's checklists. For the pilot study, data was collected and psychometric analysis was performed on the OSCE scores of 73 (seventy-three) third-year clinical medicine students in their OBGYN, Paediatrics, Internal Medicine and Surgery end of semester examination, with 31 OSCE stations altogether in June 2015. There were 11 examiners, but no standardised patients. Total number of data collected and analysed was 1690. No changes were implemented following the pilot study.

3.3.5 Data gathering

The objective of data collection is to produce reliable data (Lewis & Wojcik 2009:3 of 13). In the main study, the data were collected on the day of the OSCEs. The dependent variable in this study was all the students' OSCE scores. The independent variables were categorical in nature and included the station-examiner pair, academic levels of the examiners, the clinical subjects and gender of the students. Prior to the day of the OSCE, all the potential final year students for the July 2015 final examinations were booked for the study. The list of these students was obtained from the office of the Dean. Also, the examiners and standardised patients invited by the faculty for the July 2015 OSCE were included in the study. All these information was obtained from the office of the Dean.

On the day of the OSCEs, and before it commenced, the researcher briefed the examiners and students about the study, obtained their consent to be observed throughout the OSCEs and their scores to be used after the exercise, on the hospital grounds. The briefings lasted 30 minutes for each group of participants, on each day of the OSCE. There were no standardised patients for the July 2015 OSCE, because according to the heads of departments, there was very limited time to recruit and train willing individuals to act as standardised patients. The researcher then distributed the checklists to the research assistants and examiners as appropriate.

The OSCEs began at 09.00hrs. The live OSCE set-up (station design and station contents) and proceedings were observed and briefly described by the researcher and assistants during the OSCEs using a checklist (cf. Appendix EI). The examiners observed and assessed the students' performance in the manned stations using a clinical checklist from the Ministry of Health, Tanzania. This checklist did not cover global grades, hence, the examiners did not record accurately their global grades despite the pre-OSCE briefings. Hence, the analysis of the relationship between the global scores and the checklist scores were not done. The detailed checklist scores per examiner per candidate per station were obtained from the head of departments after marking the written stations and compilation of scores from the examiners' checklists in the manned stations.

The psychometric analysis was carried out on the scores obtained from each station, examiner and candidate as well as overall students' performance. The psychometric methods used in this study were described in Chapter 2. Hence, the dependent variable under study, to be measured, was the reliability of the OSCE scores, achieved by the students, as recorded by the examiners in the checklists. The variable (OSCE scores) is quantitative continuous in nature. The independent variables were the facets and characteristics of the testers, tests and testees operating during the OSCEs. The independent variables are categorical in nature.

The post-OSCE scores were subjected to statistical analysis in order to determine the reliability (indirectly the validity) of the scores. The statistical analysis carried out in this study were discussed in Chapter 2 and summarised in sub-section 3.4.5 below under data analysis. Statistical and text analysis were carried out using Microsoft Excel and SPSS (version 17) computer packages (Creswell 2014:16). Information gathered from literature and documentations, together with the observations and findings from the psychometric analysis of the OSCE used to test the final year clinical medicine students were used to formulate post-examination remediation and recommendations for the improvement of the OSCE at the School of Health Sciences, KIU-Dar and these can be extended to other institutions with constrained resources.

3.3.6 Data analysis

There are two main categories of statistics, namely descriptive and inferential statistics (Nick 2007:33). Both types of statistics were utilized to answer the research questions in this study. All statistical analyses in the present study were computed by using the Microsoft Excel and the SPSS statistical package for Windows version 17. "SPSS" indicates Statistical Package for the Social Sciences. The OSCE stations and scores were sorted according to the blueprint and researcher's checklists. Erroneous or missing data were excluded from the analysis and the data restructured from a multi-variate to a uni-variate design after entering into the statistical programme. A private statistician was contracted to assist in the analysis.

3.3.6.1 *Descriptive statistics*

Descriptive statistics describe the general characteristics of a distribution of scores (Salkind 2000:150; Pérez-Vicente & Expósito 2009:314). The socio-demographic characteristics of the examiners and students- participants and general information of the stations from the analysis of the researchers' observation checklists were first described. Simple descriptive statistics calculated for the study include the frequency distribution, measures of central tendencies and variation. The details of the descriptive statistics in each department include:

- Socio-demographic characteristics of examiners (gender, discipline, nationality & qualifications) and students (gender, nationality, entry qualification and year of entry).
- Analysis of the information in the researcher's checklist includes the items described in section 3.3.2.1. and also the following: number of students, number of examiners, number of manned stations, number of written stations, total number of tasks, maximum station score, station length, total duration, actual number of stations and expected number of stations.
- Frequency distribution: number of tasks per station, skewness, kurtosis, outliers and Z-scores.
- Measures of central tendencies: sum, mean, median, mode, standard setting for pass mark'.

- Measures of variability: standard deviation, range, standard error of the mean, 95% confidence interval of the mean, coefficient of variation in percentage, homogeneity of the tasks across all the stations, ANOVA and generalizability studies.

3.3.6.2 *Inferential statistics*

Breakwell *et al.* (2000:352) and Nerurkar (2008:691) define inferential statistics as the area of statistics which extends the information extracted from a sample to the actual environment in which the problem arises. It therefore seeks to draw inferences from the sample about the population and to test hypothesis.

Parametric vs. non-parametric statistics

Inferential statistics is divided into two broad categories, namely parametric and nonparametric statistics (Otwombe, Petzold, Martinson & Chirwa 2014:3 of 10). Non-parametric statistical analysis requires few assumptions about the population and focuses on the order or ranking of scores (or merely the classification function of numbers) and ignores the properties of numbers at interval and ratio scales. On the other hand, parametric tests require variables to be measured on an interval or ratio scale. Qualls, Pallin and Schuur (2014:13 of 22) explain that many studies have shown that several parametric and non-parametric tests often yield similar results. This study employed parametric tests because the dependent variable is measured on ratio scale. The parametric inferential tests of interest in this study include:

- Station analysis: t-test to determine if the generated pass mark (by standard setting) for the station was significantly different from the fixed university pass mark, difficulty index (IDI), discrimination index *(d)*, statistical significance to determine any significant difference in the means of those who passed and those who failed in each station and failure rates.
- Reliability Checks: overall alpha, alpha-correlation with overall, alpha coefficient (if-item-deleted), Pearson's correlation (*r)*, and importance of station (r^2).
- Identifying Hawks and Doves: ANOVA to identify stations with significantly high or low means compared to the rest, coefficient of variation and correlation with other stations.
- Results, reports, interpretation, discussion and recommendations (Chapters 5 and 6).
- Guideline for psychometric analysis (cf. Chapter 6).

Description of the key parametric tests used in this study is as follows:

The t-test is used to measure the significance of the difference between two means based on two independent, unrelated groups (Salkind 2000:173; Jebakumar & Manoj 2012:73). "Independence" means that the results in one group are not influenced by the results in the other group (Kim 2015:540). The t-value will be positive if the first mean is larger than the second and negative if it is smaller. However, the t-value is seldom interpreted and the p-value is used to gain an indication of the significance of the result (The Web Centre for Social Research Methods 2004:1 of 1).

Chi-square is a statistical significance test based on expected frequencies. Expected frequencies are frequencies expected in a contingency table (cross tabulation of categorical data), if the null hypothesis (no relationship) of independence were true (Jebakumar & Manoj 2012:75). Chi-square test is the basis of determining the homogeneity of the tasks across the stations in this study.

Statistical significance tests begin with the supposition that the null hypothesis is true, in other words that there is not really an effect on the population or no relationship between the variables under study. According to Breakwell *et al*. (2000:357) and Kim (2015:545), the null hypothesis simply states that there is no difference between or no real relationship exists between variables in the population. Whereas, the alternative hypothesis states that there is a real difference or relationship. If the results from the study does not agree with the initial assumption (that the null hypothesis is true), then the researcher rejects the null hypothesis and concludes that there is support for the alternative (Banerjee, Chitnis, Jadhav, Bhawalkar & Chaudhury 2009:12 of 20). Inferential statistics are therefore used to calculate the probability of obtaining the observed data if the null hypothesis is true. If the probability is small, it is unlikely that the null hypothesis is true and one could therefore conclude that the null hypothesis is false. The arbitrary cut-off point that is decided upon is called the alpha level or p-value. There is always a chance that the researcher might be wrong in his or her decision, using the probability guidelines. If the researcher rejects the null hypothesis and concludes that the population means are not equal, when in fact they are in the real population, then a Type I error was made. The reverse is said for Type II error. However, Type one error is the most commonly made error of the two. Biau, Jolles and Porcher (2010:888) explained that if the significance value is set at 0.05, it indicates that this type of error will occur 5% of the time. The most frequently used level of statistical significance, by convention is 0.05 (Breakwell *et al.* 2000:360; Banerjee *et al*. 2009:12 of 20). For some studies on particularly controversial topics or where making a type I error could have critical consequences, a more strict level could be chosen. For the purpose of this study however, the significance level of 0.05 is considered adequate (Castleman 2007:101).

Hypothesis

The null hypothesis being tested here is that the limited resources of medical schools in challenged regions of Africa as an example, have no effect on the psychometric properties of the OSCEs practiced in these schools.

The alternative assumption would be that the challenges of limited resources in these medical schools are strong enough to significantly influence the psychometric properties of the OSCEs practiced in these schools.

3.4 LIMITATIONS

One of the major resource challenges in the selected medical school for study, was that, it does not have its own teaching hospital yet as it was under construction at the time this study was carried out (cf. Figure 3.4). This might affect the smooth planning and administration of the OSCE, as the university has to comply with the hospital regulations. One of such restriction was observed in the administration of questionnaires to the real patients that participated in the study, since there were no standardised patients. The hospital administration disallowed the participation of the real patients in the study with the claim that the research will be too stressful for the real patients.

Time was also a constraint in the study. The final examinations were chosen because they were expected to be the best of all the examinations implemented in the university, in terms of the proportion of university resources invested in its preparation and implementation and is also the examination that will produce the graduates. All eyes are on this set of examinations, as it is a measure of the quality of the graduates, who sat for it. However, this examination is restricted to once a year in July and therefore the researcher must organise her schedule around this exam. Once, missed, the next opportunity, will be a year after.

The researcher did not have access to the completed examiners' checklists as it was strictly confidential. Hence, the researcher relied on the marks compiled by the Heads of department for analysis. Detailed analysis of the content of the checklist could not be carried out in this study. Despite such limitations faced, the study findings are authentic and can be relied on to inform the revision of the management of OSCE.

3.5 ASSUMPTION

The two assumptions held by the researcher and statistician for statistical analysis were that:

- The students' scores in each station followed at least an approximately normal distribution.
- The stations' scores are independent of each other. That is, the value of one score is not related with the value of another.

(Kuzma & Bohnenblust 2001:161)

3.6 ENSURING QUALITY

The following measures were taken to ensure quality in the current study.

3.6.1 Validity

Validity as defined earlier, is the extent to which an instrument measures what it purports to measure or how truthful the research results are (Kimberlin & Winterstein 2009:2276). By adhering to the approved study protocol, choosing appropriate evaluation instruments that are specifically recognized by the AMEE (Association of Medical Educators in Europe), reputable journals, institutions and authors, have been repeatedly used and subjected regularly to peer review, to assess the components of interest in this study, the validity of this study results was enhanced. Moreover, piloting the data collection tools and involving more than one observer in the study has also increased its validity.

3.6.2 Reliability

Reliability is defined as the extent to which a questionnaire, test, observation or any measurement procedure produces the same results on repeated trials (Heale & Twycross 2015:1 of 2). Piloting and use of standard data collection tools improved the reliability in this study.

3.6.3 Trustworthiness

Trustworthiness is the demonstration that the evidence for and the arguments for the research results reported are sound and strong. According to Labanca (2010:1 of 1) and Noble and Smith (2015:34) described the criteria for trustworthiness in quantitative research, namely: internal and external validity, reliability, generalizability and objectivity. The literature review of previous, similar studies done elsewhere, piloting, use of structured observation, team work with assistants and consultation with a registered private statistician, use of various psychometric methods, peer debriefing by my two promoters and feedbacks from several seminar and conference presentations increased the trustworthiness of this study.

3.7 ETHICAL CONSIDERATIONS

The following ethical precautions were taken in this study to ensure compliance with the code of ethics in research (Mandal, Acharya & Parija 2011:2).

3.7.1 Approval

Approvals for the research project were obtained from the ethics committee of the Faculty of Health Sciences at the University of the Free State, where the study originated from; the Dean of the Faculty of Health Sciences at the University of the Free State, the management of Kampala International University, Dar es Salaam campus and the management of the Amana regional referral hospital, Dar es Salaam, which was the study site.

3.7.2 Informed consent

Written informed consents were obtained from the examiners. As earlier mentioned, there were no standardised patients for this July 2015 OSCE, hence, the questionnaires were not served. Group and verbal consent was obtained from the participating students, to allow themselves to be observed by the researcher and assistants. The participating examiners consented to be observed during the OSCE and the scores they awarded to the examinees during the OSCE to be subjected to psychometric analysis. The OSCE processes were observed and the researcher assured the university and hospital that the observation checklists and the OSCE scores will be handled with high level of confidentiality, anonymous and security. A short overview of the study and its purposes were provided to the participants with an explanation of what is required from them. The researcher’s name and contacts were made available to the participants, who would have access to the results of this study.

3.7.3 Right to privacy

Number coding was used to ensure confidentiality, where necessary. No names or personal identifiers appeared on any data score sheet that were sent for statistical analysis. All information was managed in a strictly professional and confidential manner.

3.8 CONCLUSION

Chapter 3 provided an overview of the research design methodology involved in the study and the procedures that were followed. Piloting to test the data collecting tools and the Psychometric methods was also successfully carried out and reported in this chapter.

In Chapter 4, **Report of the results of the main study** will be presented.

CHAPTER 4

RESULTS OF THE PSYCHOMETRIC ANALYSIS OF THE OSCE's

4.1 INTRODUCTION

The purpose of this chapter is to present the results of this study. Data for this study was collected in four clinical departments in which final year students carried out their OSCE between 27th - 31st July 2015. All statistical analyses in the present study were computed using SPSS statistical package for Windows version 17 and Microsoft Excel. For all the statistical data, the percentages are rounded off to one decimal place. Recall that this study aimed to achieve the following objectives:

- Describe the quality of the OSCE that is currently practiced in a typical resource limited medical school at KIU-Dar in Dar es Salaam. Illustrate the application of the available psychometric methods on the raw scores from a live summative OSCE in a resource limited medical school at KIU-Dar in Dar es Salaam to obtain its psychometric properties and provide an appropriate interpretation of these properties.
- Describe the available psychometric methods for the OSCE through literature review.
- Develop a guideline for psychometric analysis and interpretation of the OSCE scores in a medical school. Describe how the findings of the analysis and its interpretation can be used to improve the OSCE at KIU-Dar in Dar es Salaam over a period of time.

Hereunder, the findings in answer to the above objectives are shown and discussed. The OSCE took place, off-campus in Amana regional referral hospital, Dar es Salaam because the Kampala International University-Dar es Salaam campus was yet to have its own teaching hospital. The four clinical departments involved in the July 2015 final OSCE's were OBGYN, Paediatrics, Internal Medicine and Surgery. Of note is that the Psychiatry OSCE's was covered in Internal Medicine. There was a total of thirty (30) clinical students booked for the OSCE's. For the purpose of the study, each candidate was given an identification number (ID) as it appeared on the faculty register in the Dean's office and this same number was consistently used to track each of the students throughout the OSCE's, Table 4.1. Permissions to carry out the study were already obtained from the HODs, university and hospital administration.

The main aim of the data analysis was to describe the socio-demographic features of the candidates and examiners as well as the OSCE structure and proceeding. A report on the analysis of the stations, subjects and overall metrics was also presented. The other attributes of the OSCE's, such as its stability with regards to the correlation between the manned and written stations' scores, the candidate gender and their means, academic levels of participating examiners and the G-coefficients were also discussed. Lastly, in this chapter, the quality properties of the borderline method that was used to determine the pass mark was analysed.

Psychometric analysis was carried out at station, subject (total station metrics) and faculty (overall station metrics) levels as described in Chapter 2 (cf. 2.5.4), Chapter 3 (cf. 3.3.6) and summarized below:

Descriptive Statistics

- Socio-demographic characteristics of examiners (gender, discipline, nationality and qualifications) and students (gender, nationality, entry qualification and year of entry).
- Analyses of the information in the researcher's checklist include the items described in section 3.2.2.2 and also the following: number of students, number of examiners, number of manned stations, number of written stations, total number of tasks, maximum station score, station length, total of duration, actual number of stations and expected number of stations.
- Frequency distribution: number of tasks per station, skewness, kurtosis, outliers and Z-scores.
- Measures of central tendencies: sum, mean, median, mode, standard setting for pass mark.
- Measures of variability: standard deviation, range, standard error of the mean, 95% confidence interval of the mean, coefficient of variation in percentage, homogeneity of the tasks across all the stations, ANOVA and generalizability studies.

Inferential Statistics

- Station analysis: t-test to determine if the generated pass mark (by standard setting) for the station was significantly different from the fixed university pass mark, difficulty index (IDI), discrimination index *(d)* and statistical significance to determine any significant difference in the means of those who passed and those who failed in each station and failure rates.
- Reliability checks: overall alpha, alpha-correlation with overall, alpha coefficient (if-item-deleted), Pearson's correlation (*r)* and importance of station (r^2).

- Identifying Hawks and Doves: ANOVA to identify stations with significantly high or low means compared to the rest, coefficient of variation and correlation with other stations.

4.2 REPORT ON THE OSCE OBSERVATIONS, CHECK LIST AND QUESTIONNAIRE

The OSCE design and proceedings were observed and described by the researcher and two assistants in unison. Day 1-4 OSCE's were in this order: OBGYN, Paediatrics, Internal Medicine and Surgery. Their findings were recorded with checklists (cf. section 3.2.2.2., Appendix EI) and summarised in Table 4.1 below.

On each day of the OSCE, the examiners and students arrived at the referral hospital at 07.00hours. The students were conveyed to and from the hospital by the university bus. The examiners arrived early, except in OBGYN and Medicine where the majority of the examiners were part-timers (cf. Table 4.3). After breakfast, the examiners and students were briefed by the programme coordinator and researcher for 45 minutes and provided their consent for the research. The programme coordinator served the examiners, the checklists and the students were ushered to the different hospital wards according to the exam schedule. Two of the examiners in OBGYN and one in medicine missed the pre-OSCE briefing. The research assistants briefed them about the study. The OSCE's began at 09.00hours in all the departments except in OBGYN, where the students experienced about 1 hour delay before the OSCE commenced. There were no standardised patients, but real patients were used for the OSCE's. The selection of the real patients was the responsibility of each of the hospital departmental heads. The very ill patients were excluded from the exercise. The real patients and their care-givers were cooperative. The patients were rewarded with refreshments for their cooperation during and at the end of the OSCE.

4.2.1 OSCE observations

Twenty-seven out of the 30 students in the faculty register sat for the July 2015 final OSCE and were assessed by 20 examiners in 82 stations (cf. Table 4.1). The expected total number of OSCE stations was 80 (twenty for each subject). There were 10 students for OBGYN, 19 for Paediatrics, 23 for Medicine and 19 for Surgery. The examiners were 5, 4, 5 and 6 in number, in the same subject order. OBGYN had a total of 4 stations, Paediatrics-27, Medicine-26 and Surgery-25. The OBGYN department could not generate sufficient number of OSCE stations due to lack of stable human resource, since all the academic staff in the department were part-timers (cf. Table 4.3). There were altogether 17 stations manned by one examiner each including the senior assessors. OBGYN had 2, Paediatrics had 4, Medicine had 5 and Surgery had 6 manned stations. The rest of the stations were in written format. The number of tasks in each station was between 1-6 and the students attempted a total of 172 tasks, divided as follows: 10 in OBGYN, 56 in Paediatrics, 73 in Medicine and 33 in Surgery (cf. Table 4.1). The maximum score for each station was 10 marks and the length of each station was 5 minutes. The July 2015 OSCE was conducted in a total of 470 minutes, one subject per day, over 4 days. OBGYN OSCE lasted for 30 minutes and the OSCE duration for each of the remaining subjects was 150 minutes (cf. Table 4.1).

The manned stations were designed around the selected real patients' beds and the unmanned or written stations were set up on tables in the centre of each ward. The manned stations used in the OSCE were of different categories as shown in Table 4.1 below. The 2 manned stations in OBGYN and the 5 in Medicine assessed only history taking and physical examination skills; the 4 in Paediatrics assessed history taking, physical examination, diagnosis and procedural skills; while the 6 in surgery tested on history taking, physical examination and procedural skills. The manned stations were managed by the senior assessors (consultants and specialists) as far as possible, while the junior examiners (medical officers) marked the written stations after the OSCE. OBGYN manned stations covered only the genito-urinary system and the general physical examination, while Paediatrics and Medicine covered general physical examination, respiratory, cardiovascular and gastrointestinal systems. Surgery tested on the respiratory, gastro-intestinal, musculoskeletal, genito-urinary systems and the general physical examination. The tasks in each of the stations generally did not concentrate on just one system, rather on a blend of these system-areas. The manned stations excluded the central nervous system and the neonate. For the ease of analysis and quick reminder, the researcher moved the manned stations from their original positions to the lower end of the list of stations in each subject. For example, the manned stations in OBGYN were transferred to stations 3 and 4, Paediatrics to stations 24-27, Medicine to stations 22-26 and Surgery to stations 20-25.

The tools in the stations followed the requirements of the checklist design in all the stations. Station requirements and weightings were met according to the tasks allocated for each station and the checklists. The examiners behaved appropriately during the OSCE except in medicine, where excessive prompting of students by one of the examiners was noticed. The particular examiner was cautioned by the programme coordinator. The scores for both the manned and written stations were compiled on excel sheets by the university HODs.

An Excel copy of the OSCE results from each department was given to the researcher after a week.

TABLE 4.1: OSCE OBSERVATIONS
(Table continue on next page)

ITEM	OBGYN	PAEDS	MEDICINE	SURGERY
General Information				
Number of students (Total=27)	10	19	23	19
Number of examiners (Total=20)	5	4	5	6
Number of manned stations (Total=17)	2	4	5	6
Number of written stations (Total=65)	2	23	21	19
Number of Tasks (Total=172)	10	56	73	33
Maximum station score	10 marks			
Station length	5 minutes			
Total duration (470 minutes)	30 minutes	150minutes each		
Expected number of stations (Total=80)	20	20	20	20
Type of manned stations	**OBGYN**	**PAEDS**	**MEDICINE**	**SURGERY**
i. History	station 4	station 27	station 26	station 25
ii. Physical examination	station 3	station 26	stations 22-25	stations 21-24
iii. Diagnosis	0	station 25	0	0
iv. Interpretation of Laboratory results	0	0	0	0
v. Procedural	0	station 24	0	station 20
vi. Treatment	0	0	0	0
vii. Communication/counselling skills.	0	0	0	0
Systems covered in the manned stations	**OBGYN**	**PAEDS**	**MEDICINE**	**SURGERY**
1. Central Nervous System	No	No	No	No
2. Cardiovascular System	No	Yes	Yes	No
3. Respiratory system	No	Yes	Yes	Yes
4. Gastrointestinal system	No	Yes	Yes	Yes
5. Musculoskeletal system	No	No	No	Yes
6. Neonatal	No	No	No	No
7. Genitourinary system	Yes	No	No	Yes
8. General Examination	Yes	Yes	Yes	Yes
Tools in the stations follow the requirements of the checklist design	**OBGYN**	**PAEDS**	**MEDICINE**	**SURGERY**
1. Strongly agree				
2. Agree	🟩	🟩	🟩	🟩
3. Neutral				
4. Disagree				
5. Strongly disagree				

All assessors arrived early	**OBGYN**	**PAEDS**	**MEDICINE**	**SURGERY**
1. Strongly agree		■		
2. Agree				■
3. Neutral			■	
4. Disagree	■			
5. Strongly disagree				
Assessors were briefed	**OBGYN**	**PAEDS**	**MEDICINE**	**SURGERY**
Yes	■	■	■	■
No				
Students were briefed	**OBGYN**	**PAEDS**	**MEDICINE**	**SURGERY**
Yes	■	■	■	■
No				
Behaviour of assessors were appropriate	**OBGYN**	**PAEDS**	**MEDICINE**	**SURGERY**
1. Strongly agree				
2. Agree	■	■		■
3. Neutral			■	
4. Disagree				
5. Strongly disagree				

4.2.2 Examiner's checklist and standardised patient's questionnaire

The examiner's checklist was described in section 3.2.2.2 and Appendix EII. The assessor's checklist, which was supplied by the Ministry of Health in Tanzania, did not cover global grades. Hence, the examiners were not consistent in entering the global scores for each student because the majority of them did not understand the concept of global grading despite the initial briefing. The few examiners that entered their global scores, directly translated the checklist scores to global grades, the act of which was incorrect. Moreover, the researcher did not have access to the completed examiners' checklists and therefore detailed analysis of the checklist could not be carried out.

There was no standardised patient during the July 2015 final OSCE's. Accordingly, no questionnaire was served. Real patients were used for the OSCE's, but the hospital authorities did not permit the researcher to administer questionnaires to the real patients nor interview them for the purpose of research. They claimed that 'the patients will be stressed too much'.

4.3 SOCIO-DEMOGRAPHIC DESCRIPTION OF THE STUDENTS

There were 30 students altogether, 26 (86.7%) male and 4 (13.3%) female in the final year. Only 27 (23 male and 4 female) out of these 30 candidates sat for the July 2015 final OSCE. The three candidates (ID 7, 28 and 29, all male) that did not sit for this OSCE in any of the departments, had not cleared their tuition fees as at the time of final examination. The 27 candidates included 14 (52%) regular students, 11 supplementing students and 2 repeating students. The regular students sat for the four subjects except candidates' number 24-27, who were not allowed to sit for the OSCE's of the theory paper they failed. Of the regular students, candidates' number 24 and 25 did not sit for Paediatrics OSCE, ID number 26 did not sit for OBGYN OSCE, while ID number 27 did not sit for paediatrics and OBGYN OSCE's. The regular students for this 2015/2016 final year class were students with the 305 and 304 university registration numbers, who had been progressing well. According to the rule of the faculty, only students that pass the theory paper are allowed to sit for the OSCE of that subject. Two candidates (ID 21 & 22) with 202- and 201- university registration numbers were repeating that class. The remaining 11 candidates were supplementing in 1-3 subjects. All these students were admitted through direct entry with ordinary secondary school certificate, except for the two repeating students of the 201/202 series (ID 21 & 22). These repeating students were admitted with advanced secondary school certificate. All the 30 students were Tanzanians by nationality.

Ten (7 male and 3 female candidates) out of the 30 students in the faculty register sat for the OBGYN final OSCE in July 2015 and were assessed by 5 examiners in 4 stations. Nineteen (16 male and 3 female candidates) sat for the Paediatrics final OSCE and were assessed by 4 examiners in 27 stations. Twenty-three (20 male and 3 female students) sat for Internal Medicine OSCE and were assessed by 5 examiners in 26 stations. Nineteen (15 male and 4 female candidates) sat for the surgery final OSCE and were assessed by 6 examiners in 25 stations (cf. Table 4.2).

TABLE 4.2: SOCIO-DEMOGRAPHIC DESCRIPTION OF THE STUDENTS

	GENDER		ACADEMIC STATUS			ENT QUALIFICATION		SUBJECTS			
ID	MALE	FEMALE	REGULAR	SUPP	REPEAT	O LEVEL	A LEVEL	OBGYN	PEADS	MED	SURG
1.00	■			■		■			.	■	■
2.00	■			■		■			■	■	
3.00	■			■		■			■	■	
4.00	■			■		■			■	■	
5.00	■		■			■		■	■	■	■
6.00	■		■			■		■	■	■	■
7.00	■					■					
8.00	■		■			■		■	■	■	■
9.00	■		■			■		■	■	■	■
10.00		■	■			■		■	■	■	■
11.00	■		■			■		■	■	■	■
12.00	■			■		■			■		
13.00	■			■		■			■		
14.00	■		■			■		■	■	■	■
15.00	■			■		■					■
16.00		■	■			■		■	■	■	■
17.00	■		■			■		■	■	■	■
18.00	■		■			■		■	■	■	■
19.00	■			■		■				■	
20.00	■			■		■				■	■
21.00	■				■		■	■	■	■	■
22.00	■				■		■	■	■	■	■
23.00	■			■		■			■		
24.00		■	■			■		■		■	■
25.00	■		■			■		■		■	■
26.00		■	■			■			■	■	■
27.00	■		■			■				■	■
28.00	■					■					
29.00	■					■					
30.00	■			■		■				■	
Total	26	4	14	11	2	28	2	14	19	23	19

4.4 SOCIO-DEMOGRAPHIC DESCRIPTION OF THE EXAMINERS

There were 20 examiners altogether (5 in Internal Medicine, 5 in OBGYN, 4 in Paediatrics and 6 in Surgery). Amongst these examiners, 4 were female and 16 male. There was no female examiner in Surgery. Tanzanian examiners were 15 (2 Ugandans, 2 Kenyans and

1 Nigerian). The invited external examiners (consultants) were 4 in number and they were not staff of KIU-Dar. Four of the internal examiners were fulltime staff, while 12 were part-time staff of KIU-Dar. Eight of the examiners had MMed (Masters) qualification and the rest (12) were first-degree holders (MBCHB-Medical Officers). There was no examiner with PhD qualification. Each department had one consultant as external examiner and the rest of the MMed Holders were specialists. All the internal examiners in OBGYN were part-timers (cf. Table 4.3).

TABLE 4.3: SOCIO-DEMOGRAPHIC DESCRIPTION OF THE EXAMINERS

	DEPT	OBGYN		PAEDS		MEDICINE		SURGERY	
		MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
NATIONALITY	TANZANIAN	4	1	1	1	3	1	4	0
	UGANDAN	0	0	0	0	1	0	1	0
	KENYAN	0	0	1	0	0	0	1	0
	NIGERIAN	0	0	0	1	0	0	0	0
EMPLOYMENT STATE	FULLTIME	0	0	1	1	1	0	1	0
	PARTTIME	4	0	0	1	2	1	4	0
QUALIFICATIONS	MBCHB	3	0	1	1	1	1	5	0
	MMED	1	0	0	1	2	0	0	0
	CONSULTANT (MMED): EXTERNAL EXAMINER	0	1	1	0	1	0	1	0
TOTAL EXAMINERS	**20**	**5**		**4**		**5**		**6**	

4.5 REPORT ON STATION METRICS IN OBGYN

The report on station metrics in OBGYN is given in Section 4.5.1 below.

4.5.1 OBGYN stations metrics

The manned stations had two tasks each while the written stations had three tasks each (cf. Table 4.4) below. The following are the metrics generated from the OBGYN stations.

4.5.1.1 *Frequency distribution*

Skewness: All the stations' score distribution was skewed (cf. Test 2 in Table 4.4, Figure 4.1). Stations 1-3 were negatively skewed. Station 4 and the total OBGYN scores had 0.1 and 0.5 skewness respectively. The lowest negative skewness was in station 3 with -1.8.

TABLE 4.4: OBGYN OSCE STATIONS SCORES DISTRIBUTION

TEST	STATION 1	STATION 2	STATION 3	STATION 4	TOTAL
1. Number of Tasks	3	3	2	2	10
2. Skewness	-1	-1. 8	-1.8	0.1	0.5
3. Kurtosis	2.3	1.4	3.4	-1	-0.8
4. Outliers	Bottom	Bottom	Bottom	—	—

Kurtosis: Stations 1-3 have positive kurtosis. The kurtosis of station 3 is pronounced. Station 4 and the OBGYN total scores have negative kurtosis, which is lower in station 4 at -1, (cf. Test 3 in Table 4.4, Figure 4.1).

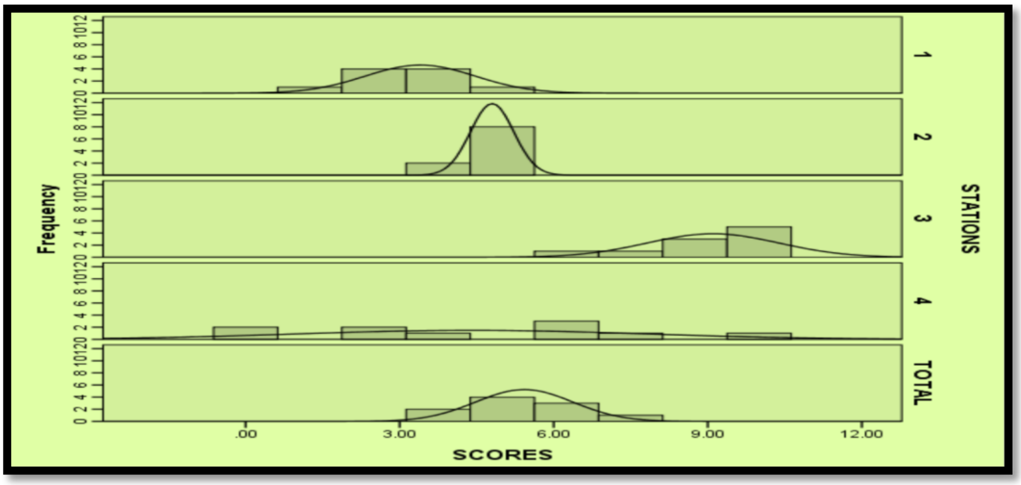

FIGURE 4.1: OSCE SCORES DISTRIBUTION IN OBGYN STATIONS

Checking for outliers with Box-whiskers plot

According to Test 4 in Table 4.4, stations 1-3 have low extreme scores (cf. Figure 4.2).

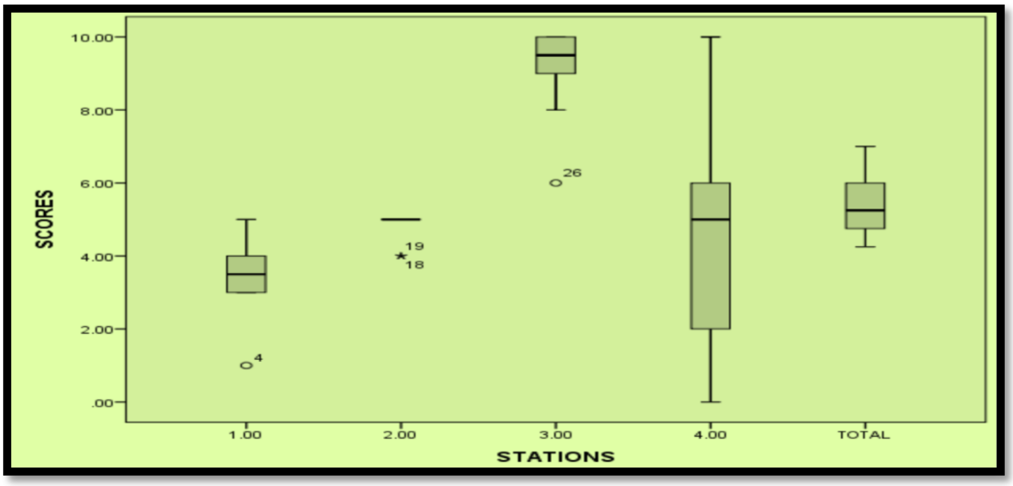

FIGURE 4.2: CHECKING FOR OUTLIERS IN OBGYN STATIONS

Z-scores

After standardizing the raw scores, the Z-scores were grouped into 7 grades (A-G) as follows: Grade A [excellent: >+3 (0.13% of the data in a normal data distribution)]; Grade B [Very good: ≥+2 ≤ +3 (2.15% of the data in a normal data distribution)]; Grade C [Good: >+1<+2 (13.59% of the data in a normal data distribution)]; Grade D [Average: +1≤0≥-1 (68.26% of a normal data distribution)]; Grade E [Poor: +<-1>-2 (13.59% of a normal data distribution)]; Grade F [Very Poor: -2≥-3 (2.15% of a normal data distribution)]; Grade G [Fail: <-3(0.13% of a normal data distribution)]. The frequencies of the candidates' performances in each group were compiled below in Table 4.5. There were no extreme values in the normalised scores. The Z-letter grades corresponded with different scores in each station. The best performance in OBGYN was good (grade C) and the worst was poor (grade E).

TABLE 4.5: SUMMARY OF OBGYN Z-SCORES

GRADES	GOOD	AVERAGE	POOR	TOTAL
Grades	C	D	E	
Z-Score ranges (%)	>+1<+2 (13.6)	+1≤0≥-1 (68.3)	<-1>-2 (13.6)	100%
Frequency	2 (20%)	6 (60%)	2 (20%)	10 (100%)
Raw Score Equivalent	6.5-6.9	4.8-5.9	4-4.2	

4.5.1.2 *Measures of central tendencies*

All the measures of central tendencies (sum, mean, mode, median and standardised pass marks) in the OBGYN stations were highest in station 3 and lowest in station 1, (cf. Tests 5,6,7,8 & 9 in Table 4.6). The mean, median and mode were not equal in any of the stations. However, these measures are in proximity to each other except in station 4. The standardised pass mark in station 2 was the same as the university pass mark of 5, while stations 1 and 4 standardised pass marks were below 5. The standardised pass marks in station 3 and the total were above 5, especially in station 3.

TABLE 4.6: OBGYN OSCE STATIONS: MEASURES OF CENTRAL TENDENCIES

TEST	STATION 1	STATION 2	STATION 3	STATION 4	TOTAL
5. Sum	34	48	91	44	54.3
6. Mean	3.4	4.8	9.1	4.4	5.4
7. Median	3.5	5	9.5	5	5.3
8. Mode	3	5	10	6	4.3
9. S. Setting	3.5	5	9.5	4.5	5.3

4.5.1.3 *Measures of variability*

The range, standard deviation, standard error of the mean and coefficient of variation were widest in station 4 and narrowest in station 2 (cf. Tests 10,11,12 & 14 in Table 4.7). The station tasks were significantly heterogeneous (cf. Test 16 in Table 4.7). The 95% confidence interval of the mean in stations 1 and 3 does not contain the university pass mark of 5 (cf. Test 13 in Table 4.7).

TABLE 4.7: OBGYN OSCE STATIONS: MEASURES OF VARIABILITY

Tests	1	2	3	4	Total
10. S. Deviation	1.1	0.4	1.3	3.4	1.0
11. Range	4	1	4	10	2.8
12. SEM	0.3	0.1	0.4	1.1	0.3
13. 95% CI(Mean)	2.63-4.2	4.5-5.1	8.2-10.0	2.0-6.8	4.8-6.1
14. C. Variation (%)	31.5	8.8	14.2	76.6	17.5
16. Homogeneity	L =12.2, *p*<0.05, Station tasks are heterogenous				

ANOVA (comparing means): The variance between the stations is very high (>30%). Moreover, the variance between the stations was significantly higher than the variance within the stations. The mean of station 3 was consistently and significantly higher than the rest. (cf. Test 25 in Table 4.8).

TABLE 4.8: ANOVA TABLE for OBGYN OSCE

VARIANCE	SS*	*Df***	MS***	F****	Sig*****
Between Groups	190.5 (59.6%)	3	63.5	17.7	0.0
Within Groups	129.3 (40.4%)	36	3.6		
Total	**319.8 (100%)**	**39**			
	Turkey's HSD Post Hoc test= Station 3 mean is significantly higher than the rest.				

SS*: Sum of squares; Df** Degrees of freedom; MS***: Mean of Squares; F****: Ratio of between mean of squares and within mean of squares; SS*****: Significance.

Generalizability studies: variance components estimates

The students and the examiners contributed 0.1% and 89.2% respectively to the variance obtained in the OBGYN OSCE. The interaction between students and examiners was high at 10.7%. G-coefficient was low at 0.0 (cf. Table 4.9). The level of examiners' errors and student-examiner interactions experienced in the OBGYN stations is significantly high enough to influence the OSCE scores.

TABLE 4.9: GENERALIZABILITY STUDIES FOR OBGYN OSCE

GENERALIZABILITY STUDIES: VARIANCE COMPONENTS ESTIMATES		
	Components	**%**
Students	0.0	0.1%
Examiners	6	89.2%
Students*Examiners	3.6/5 examiners	10.7%
Total	**6.7**	
G-coefficient= 0.0/0.7=0.0. Errors from examiners was 89.2%		

4.5.1.4 *Station analysis*

T-test: The standardised pass marks in stations 1 and 3 were significantly different from the university pass mark of 5 (cf. Test 15 in Table 4.10).

Item Difficulty Index (IDI): Good IDI is between 0.3 and 0.8, furthermore, for a criterion referenced test, IDI of 0.9 is acceptable (cf. Chapter 2, Section 2.5.4.4). Stations 2, 4 and the total OBGYN scores had IDI within the good range. Station 3 IDI was above, while station 1 IDI was below the good range (cf. Test 17 in Table 4.10).

Station Discrimination Index (*d*): The range of *d* is -1.00 to +1.00 (good range is between 0.3-0.5). Only station 2 had *d* within the good range with 0.3 (cf. Test 18 in Table 4.10). The rest of the stations had *d* outside the good range. *d* in station 1 was negative. The situation in station 1 is such that the academically strong students were failing and the weak students were passing the tasks. The tasks in station 1 should be discarded.

Statistical Significance: The mean of those who passed was significantly higher than the mean of those who failed the OBGYN OSCE in station 4 (cf. Test 19 in Table 4.10). Pass/Fail means were not significantly different in the other stations. The pass/fail means were the same in Station 3.

Failure rates: The highest failure rate was found in station 1 with 9 (90%). No one failed in station 3 (cf. Test 24 in Table 4.10).

TABLE 4.10: OBGYN OSCE STATIONS ANALYSIS

TEST	1	2	3	4	TOTAL
15. T-test(Pass mark)	T=4.4*	T=0	T=13.2*	T=1.5	T=0.9
17. IDI	0.1	0.8	1	0.5	0.7
18. *d*	-0.3	0.3	0	1	
19. S. Sig (P/F)	T=0.5, *p*=0.6	T=1.3, *p*=0.2	T=0.3, *p*=0.8	T=5.5*	
24. Failure rates	9(90%)	2(20%)	0	5(50%)	3(30%)

**P<0.05, significant*

4.5.1.5 *Reliability checks*

Alpha-correlation with overall: The total alpha coefficient for the OBGYN stations was 0.008, which was very poor. Stations 1 and 2 have negative correlations with the total alpha. Stations 3 and 4 had positive, but very poor correlations with the total alpha. Station 1 had the poorest correlation with total alpha with -0.1 (cf. Test 20 in Table 4.11).

Alpha coefficient (if-item-deleted): When each station was deleted in turn, alpha in stations 1 and 2 improved to above the total alpha, while the alpha in stations 3 and 4 decreased. (cf. Test 21 in Table 4.11).

Pearson's correlation, r (station with student total) and r^2: Pearson's correlation and contribution to the total score variance was significantly high in station 4 with 0.9 (cf. Tests 22 & 23 in Table 4.11). Correlations and contributions were low in the other stations.

TABLE 4.11: OBGYN OSCE STATIONS: RELIABILITY CHECKS

TEST	1	2	3	4	TOTAL
20. α Correlation	-0.1	-0.0	0.0	0.1	0.0
21. α Deleted	0.1	0.0	-0.0	-0.7	
22. Pearson Corr(r)	r=0.1, p=0.7	r=0.1, p=0.8	r=0.4, p=0.3	r=0.9*	
23. r^2	0.0	0.0	0.1	0.9	

**P<0.05, significant*

Guidelines for the interpretation of correlation coefficients: 0.75-1.00: strong; 0.50-0.74: moderate to high; 0.25-0.49: low to moderate; 0.00-0.24: weak.

4.5.1.6 *Identifying Hawks and Doves in OBGYN OSCE stations*

Station 3 was marked too leniently based on the following evidences:

- The scores distribution is negatively skewed (cf. Test 2 in Table 4.4, Figure 4.1)
- Has excess kurtosis (cf. Test 3 in Table 4.4, Figure 4.1).
- Has several outliers (cf. Test 4 in Table 4.4, Figure 4.2).
- Has significantly higher mean (cf. Test 25 in Table 4.8).
- Variability is small (cf. Tests 10-16 in Table 4.7).
- IDI is too high (cf. Test 17 in Table 4.10).
- Discriminating power is very poor (cf. Test 18 in Table 4.10).
- Correlations and contribution are low (cf. Test 22-23 in Table 4.11).

4.6 REPORT ON STATION METRICS IN PAEDIATRICS

The report on station metrics in Paediatrics is as follows:

4.6.1 Paediatrics stations metrics

4.6.1.1 *Frequency distribution*

Skewness: All the stations' score distribution were skewed (cf. Test 2 in Table 4.12, Figure 4.3). Eight stations (7,8,10,14,15,17,19,27 & the total) were positively skewed. Station 19 has the highest positive skewness of 3.0. The rest of the stations were negatively skewed. The lowest negative skewness was in station 26 with -1.6.

TABLE 4.12: PAEDIATRICS OSCE STATIONS: SCORES DISTRIBUTION

TESTS STATIONS	NUMBER OF TASKS (1)	SKEWNESS (2)	KURTOSIS (3)	OUTLIERS (4)
1	2	-0.2	-0.3	Bottom
2	2	-0.3	0.5	Bottom, Top
3	1	-0.2	-1.2	---
4	2	-0.6	1.0	Bottom
5	3	-0.1	0.8	--
6	1	-0.3	-0.6	--
7	1	0.2	-1.4	--
8	2	1.3	-0.3	Top
9	3	-0.0	0.1	--
10	3	0.1	-0.2	Bottom, Top
11	3	-0.5	-0.3	--
12	4	-1.3	0.9	---
13	2	-1.2	-0.4	--
14	2	1.1	0.3	--
15	2	0.1	-1.1	--
16	3	-0.7	0.4	---
17	2	0.8	-0.3	--
18	4	-0.9	1.2	Bottom
19	2	3.0	8.5	Top
20	2	-0.9	-0.1	Bottom
21	1	-0.6	0.2	--
22	1	-0.3	-0.6	--
23	2	-1.1	0.2	Bottom
24	1	-0.7	-0.8	--
25	1	-0.6	-1.2	--
26	1	-1.6	2.4	Bottom
27	3	0.1	-0.1	Top
TOTAL	**56**	**0.1**	**-1.1**	**--**

Kurtosis: Twelve stations (2,4,5,9,12,14,16,18,19,21,23 & 26) have positive kurtosis. The kurtosis (8.5) of station 19 is pronounced. The rest of the stations and the Paediatrics total scores have negative kurtosis, with the lowest in station 7 at -1.4 (cf. Test 3 in Table 4.12, Figure 4.3).

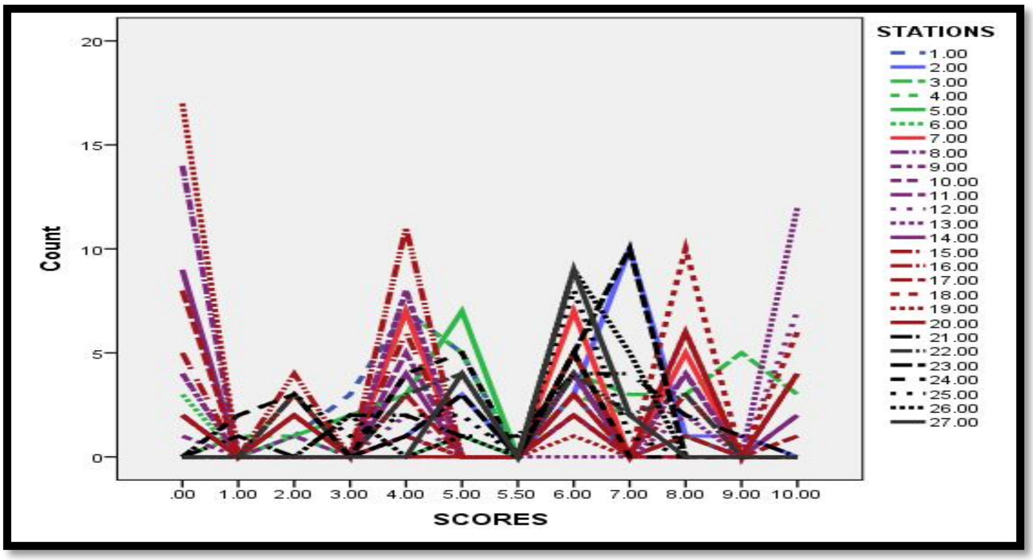

FIGURE 4.3 OSCE SCORES DISTRIBUTION IN PAEDIATRICS STATIONS

Checking for outliers with Box-whiskers plot

Six stations (1,4,18,20,23 & 26) have low extreme scores, while 3 stations (8,19 & 27) have large extreme scores and 2 stations have both large and low extreme values (Test 4 in Table 4.12, Figure 4.4).

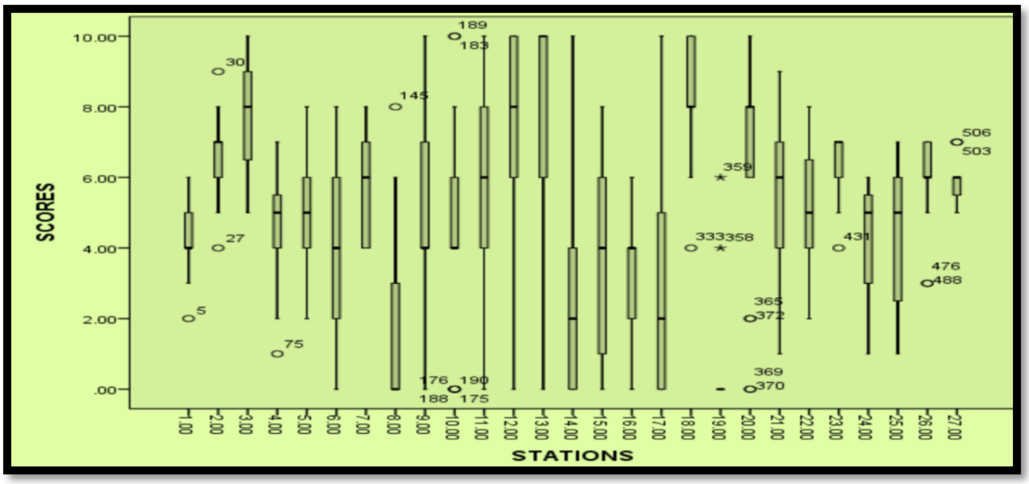

FIGURE 4.4: CHECKING FOR OUTLIERS IN PAEDIATRIC STATIONS

Z-scores (cf. Table 4.13): There were no extreme values after normalising the raw OSCE scores. The z letter grades corresponded with different scores in each station. The best performance in Paediatrics was good (grade C) and the worst was poor (grade E).

TABLE 4.13: SUMMARY OF PAEDIATRICS Z-SCORES

GRADES	GOOD	AVERAGE	POOR	TOTAL
GRADES	C	D	E	
Z-SCORE RANGES	>+1<+2 (13.6%)	+1≤0≥-1 (68.3%)	<-1>-2 (13.6%)	100%
Frequency	4(21%)	11(57.9%)	4(21%)	19(100%)
Raw Score Equivalent	5.8-6.2	4.4-5.7	3.7-4.1	

4.6.1.2 *Measures of central tendencies*

Station 19 had the lowest measures of central tendencies (sum, mean, mode, median & standardised pass marks referred to as tests 5,6,7,8 & 9 respectively), while station 18 has the highest sum and mean in the Paediatrics stations. Station 13 has the highest median of 10 while stations 12 and 13 had the highest mode of 10 (cf. Tests 5,6,7,8 & 9 in Table 4.14). The mean, median and mode were not equal in any of the stations. However, the mean and median were in close proximity to each other except in stations 8,9,13 and 20 where they were far apart. The standardised pass marks in stations 4,5,24 and 25 were the same as the university pass mark of 5, while 10 stations (1,6,8,9,10,14,15,16,17,19 & total) had pass marks below 5. The pass marks in the rest of the 5 stations (especially station 13) were above 5.

TABLE 4.14: PAEDIATRICS OSCE STATIONS: MEASURES OF CENTRAL TENDENCIES (Table continues on next page)

Tests Stations	Sum(5)	Mean (6)	Median (7)	Mode (8)	S.Setting* (9)
1	82	4.3	4	4	4
2	124	6.5	7	7	6
3	149	7.8	8	9	8
4	88	4.6	5	4	5
5	93	4.9	5	5	5
6	72	3.8	4	4	4
7	110	5.8	6	4	5.7
8	32	1.7	0	0	0
9	100	5.3	4	4	4.9
10	84	4.4	4	4	4.3
11	116	6.1	6	4	6.3
12	138	7.3	8	10	8.3
13	140	7.4	10	10	9.7
14	52	2.7	2	0	1.7
15	70	3.7	4	4	3.7
16	64	3.4	4	4	4
17	54	2.8	2	0	2
18	156	8.2	8	8	8
19	10	0.5	0	0	0
20	122	6.4	8	8	7.1
21	102	5.7	6	6	6.1
22	100	5.3	5	5	5.4

23	119	6.3	7	7	6
24	79	4.2	5	5	5
25	84	4.4	5	6	5
26	100	5.9	6	6	6
27	88	5.9	6	6	6
Total	**94.3**	**5.0**	**4.8**	**3.7**	**4.9**

4.6.1.3 *Measures of variability*

The range, standard deviation, standard error of the mean and coefficient of variation were widest in station 13 and narrowest in station 27 (cf. Tests 10,11,12 & 14 in Table 4.15). The station tasks were significantly heterogeneous (cf. Test 16, Levene statistics: 6.0, p<0.05). The 95% confidence interval of the mean in 15 stations does not contain the university pass mark of 5 (cf. Test 13 in Table 4.15). In eight (2,3,12,13,18,23,26 & 27), out of these 15 stations, the 95% CI was above 5, while in the remaining 7 stations (1,6,8,14,16,17 & 19) the 95% CI were below 5.

TABLE 4.15: PAEDIATRICS OSCE STATIONS: MEASURES OF VARIABILITY

TESTS STATIONS	S. DEVIATION (10)	RANGE (11)	SEM (12)	95% CI (13)	C. VARIATION (%) (14)
1	1.1	4	0.3	3.8-4.8	25.7
2	1.2	5	0.3	6-7	17.9
3	1.6	5	0.4	7.1-8.6	20
4	1.5	6	0.3	3.9-5.4	32.4
5	1.4	6	0.3	4.2-5.6	28
6	2.3	8	0.5	2.7-4.9	60.7
7	1.6	4	0.4	5-6.6	28
8	2.9	8	0.7	0.3-3	174.4
9	2.4	10	0.6	4.1-6.4	46
10	3.0	10	0.7	3-5.9	68.3
11	3.1	10	0.7	4.6-7.6	50.7
12	3.2	10	0.7	5.7-8.8	44.2
13	4.2	10	1.0	5.4-9.4	56.6
14	3.4	10	0.8	1.1-4.4	122.3
15	2.9	8	0.7	2.3-5.1	77.4
16	1.6	6	0.4	2.6-4.2	48.7
17	3.2	10	0.7	1.3-4.4	110.9
18	1.6	6	0.4	7.4-9	19.7
19	1.6	6	0.4	-0.3-1.3	303.8
20	3.2	10	0.7	4.9-8	50.5
21	2.1	8	0.5	4.7-6.7	36
22	1.9	6	0.4	4.3-6.2	36.9
23	0.9	3	0.2	5.8-6.7	14.9
24	1.7	5	0.4	3.3-5	41.8
25	2.0	6	0.5	3.5-5.4	44.8
26	1.2	4	0.3	5.3-6.5	20.7
27	0.6	2	0.2	5.5-6.2	10.9
Total	**0.8**	**2.5**	**0.2**	**4.6-5.3**	**15.1**

ANOVA (comparing means): The variance between the stations was significantly high, (>30%). Moreover, the variance between the station scores was significantly higher than the variance within the stations. The mean of station 18 was significantly higher than the rest (cf. Table 4.16).

TABLE 4.16: ANOVA TABLE FOR PAEDIATRICS OSCE

VARIANCE	SS	*df*	MS	F	Sig
Between Groups	1661.2(39.3%)	26	63.89	11.91	0.00
Within Groups	2569.1(60.7%)	479	5.36		
Total	**4230.3(100%)**	**505**			
	Turkey's HSD Post Hoc test= All the Station means were significantly different from each other. Stations 18 and 19 means were significantly higher and lower than the rest of the stations respectively.				

SS*: Sum of squares; Df**: Degrees of freedom; MS***: Mean of Squares; F****: Ratio of between mean of squares and within mean of squares; SS*****: Significance.

Generalizability studies: variance components estimates

The students and the examiners contributed 7.6% and 65.8% respectively to the variance obtained in the OBGYN test. The interaction between students and examiners was 26.5%. G-coefficient=0.22 (cf. Table 4.17).

TABLE 4.17: GENERALIZABILITY STUDIES FOR PAEDIATRICS OSCE

	Components	%
Students	0.36	7.6%
Examiners	3.1	65.8%
Students*Examiners	5/4 examiners	26.5%
Total	**4.71**	
	G-coefficient= 0.36/1.61=0.22. Errors from examiners was 65.8%	

4.6.1.4 *Station analysis*

T-test**:** The standardised pass marks in stations 4,5,9,24,25 and the total were not significantly different from the university pass mark of 5 (cf. Test 15 in Table 4.18). The rest of the stations had standardised pass marks, which were significantly different from the university pass mark of 5.

Item Difficulty Index (IDI): Good IDI is between 0.3 and 0.8, furthermore, for a criterion referenced test, IDI of 0.9 is acceptable (cf. Chapter 2, section 2.5.4.4.). Nineteen stations had IDI within the good range. Five stations had IDI above the upper, while 3 stations IDI were below the lower limits of the good range respectively (cf. Test 17 in Table 4.18).

Station Discrimination Index (*d*): The range of *d* is -1.0 to +1.0 (good range is between 0.3-0.5). Six stations (1,11,12,15,24 & 26) had *d* within the good range (cf. Test 18 in Table 4.18). *d* in stations 6 and 7 were negative (-0.2 & -0.4 respectively). Stations 3 and 23 IDI was 0.

Statistical Significance: The mean of those who passed was significantly different from the mean of those who failed the paediatrics OSCE in 10 stations (2,4,5,12,13,18,20,21,23 & 26; cf. Test 19 in Table 4.18). Pass/Fail means were not significantly different in the other stations.

Failure rates: The highest failure rate was found in station 19 with 18 (94.7%). No one failed in stations 3 and 27 (cf. Test 24 in Table 4.18).

TABLE 4.18: PAEDIATRICS OSCE STATIONS ANALYSIS

TESTS STATIONS	T-test (Pmk) (15)	IDI (17)	*d* (18)	S. Sig (P/F) (19)	Failure rates (%) (24)
1	-5.9*	0.4	0.4	2.3, *p=0.1*	11(57.9)
2	5.9*	1.0	0.2	2.8*	1(5.26)
3	17.7*	1.0	0.0	-0.3, p=*0.8*	0
4	0	0.5	0.6	3.3*	9(47.4)
5	0	0.7	0.6	5.3*	6(31.6)
6	-5.9*	0.3	-0.2	0.9, p=0.4	13(68.4)
7	4.2*	0.6	-0.4	1.0, p=0.4	7(36.8)
8	-29.4*	0.3	0.6	-1.3, p=0.2	14(73.7)
9	-0.8*	0.5	0.6	2.0, p=0.09	10(52.6)
10	-4.2*	0.4	0.2	-0.2, p=0.9	12(63.2)
11	7.6*	0.6	0.4	0.2, p=0.8	7(36.8)
12	19.4*	0.8	0.4	9.3*	4(21)
13	27.7*	0.7	0.6	13*	5(26.3)
14	-19.4*	0.2	0.2	1.0, p=0.3	15(78.9)
15	-7.6*	0.3	0.4	1.2, *p=0.3*	13(68.4)
16	-5.9*	0.1	0.2	0.1, p=0.9	17(89.5)
17	-17.7*	0.3	0.2	1.8, p=0.1	14(73.7)
18	17.7*	1.0	0.2	3.6*	1(5.3)
19	-29.4*	0.1	0.2	1.5, *p*=0.2	18(94.7)
20	12.6*	0.8	0.6	3.9*	4(21.1)
21	6.4*	0.7	0.6	6.5*	5(26.3)
22	2.5*	0.7	0.2	1.4, *p*=0.2	6(31.6)
23	5.9*	1.0	0.0	2.5*	1(5.3)
24	0	0.5	0.4	1.3, *p*=0.3	9(47.4)
25	0	0.6	0.6	0.8, *p*=0.4	8(42.1)
26	5.6*	0.9	0.4	4.6*	2(10.5)
27	5.3*	1.0	0.2	-1.0, *p*=0.4	0
Total	**-0.5**	**0.4**			**12(63.2)**

*P<0.05, significant

4.6.1.5 *Reliability checks*

Alpha-correlation with overall: The total alpha coefficient for the Paediatrics OSCE was 0.6, which is moderate. Stations 3, 6, 7, 14, 16 and 27 have negative correlations with the total scores**.** The rest of the stations have positive but poor correlations with the total scores. Stations 1,9,12,13,18 and 21 have positive moderate correlation with the total scores. Station 3 had the poorest correlation with total scores with -0.4 (cf. Test 20 in Table 4.19).

Alpha coefficient (if-item-deleted): When each station was deleted in turn, alpha correlations in 25 stations improved. However, alpha in 13 of these 25 stations (2,3,6,7,10,11,14,15,16,17,23,25 & 27) improved to or above the total alpha. Alpha in stations 9 and 12 decreased (cf. Test 21 in Table 4.19).

Pearson's Correlation, *r* (station with student total) and r^2: Pearson's correlations and contributions to the total score variance was negative and very poor in stations 3 and 27. Pearson's correlation and contributions to the total score variance was significantly positive and moderately high in 10 stations (1,4,5,8,9,12,13,18,20 & 21) especially in station 21 (r=0.7, $r^2$0.5) (cf. Tests 22 & 23 in Table 4.19). Correlations and contributions were low in the other stations.

TABLE 4.19: PAEDIATRICS OSCE STATIONS: RELIABILITY CHECKS
(Table continues on next page)

TESTS STATIONS	α CORRELATION (20)	α DELETED (21)	PEARSON Corr(*r*) (22)	r^2 (23)
1	0.5	0.5	0.6*	0.4
2	0.1	0.6	0.3, *p= 0.2*	0.1
3	-0.4	0.6	-0.2, *p =0.6*	0.0
4	0.4	0.5	0.5*	0.2
5	0.4	0.5	0.6*	0.4
6	-0.0	0.6	0.1, *p =0.8*	0.0
7	-0.2	0.6	0.0, *p= 1*	0.0
8	0.4	0.6	0.6*	0.4
9	0.6	0.5	0.5*	0.3
10	0.0	0.6	0.2, *p =0.4*	0.1
11	0.1	0.6	0.4, *p =0.1*	0.1
12	0.6	0.5	0.5*	0.3
13	0.5	0.5	0.6*	0.3
14	-0.2	0.7	0.1, *p =0.6*	0.0
15	0.2	0.6	0.2, *p=0. 5*	0.0
16	-0.2	0.6	0.1, *p =0.8*	0.0
17	Z	0.6	0.3, *p =0.2*	0.1
18	0.5	0.6	0.5*	0.2
19	0.2	0.6	0.3, *p=0.2*	0.1

20	0.4	0.6	0.6*	0.3
21	0.5	0.6	0.7*	0.5
22	0.2	0.6	0.2, *p =0.4*	0.1
23	0.1	0.6	0.1, *p =0.6*	0.0
24	0.3	0.6	0.4, *p =0.1*	0.2
25	0.1	0.6	0.2, *p= 0.3*	0.1
26	0.3	0.6	0.3, *p =0.3*	0.1
27	-0.0	0.6	-0.2, *p =0.4*	0.1
Total	**0.6**	**--**	**1.00**	**1.00**

**P<0.05, significant*

Guidelines for the interpretation of correlation coefficients: 0.75-1.00: strong; 0.50-0.74: moderate to high; 0.25-0.49: low to moderate; 0.00-0.24: weak.

4.6.1.6 *Identifying Hawks and Doves in Paediatrics OSCE stations*

Stations 18 and 19 were marked too leniently and too strictly respectively based on the following evidences.

- The scores distribution were skewed (cf. Test 2 in Table 4.12, Figure 4.3)
- They had excess kurtosis especially station 19 (cf. Test 3 in Table 4.12, Figure 4.3).
- There were several outliers (cf. Test 4 in Table 4.12, Figure 4.4).
- Both stations had significantly different means from the rest of the stations (cf. Test 25 in Table 4.16).
- Variability is too high (cf. Tests 10-16 in Table 4.15).
- IDIs are outside the good range (cf. Test 17 in Table 4.18).
- Discriminating powers are very poor (cf. Test 18 in Table 4.18).
- Correlations and contribution are low (cf. Tests 22-23 in Table 4.19).

4.7 REPORT ON STATION METRICS IN INTERNAL MEDICINE

The report on station metrics in Internal Medicine is as follows:

4.7.1 Internal Medicine stations metrics

4.7.1.1 *Frequency distribution*

Skewness: All the stations' score distributions were skewed (cf. Test 2 in Table 4.20, Figure 4.5). Nineteen stations (3-7,9,10,13,16,17,20,22-24,26 & total) were negatively skewed. The lowest negative skewness was in station 5 with -2.22. The rest of the stations were positively skewed, highest in station 15 (1.8).

TABLE 4.20: INTERNAL MEDICINE OSCE STATIONS SCORES DISTRIBUTION

TESTS STATIONS	NUMBER OF TASKS (1)	SKEWNESS (2)	Kurtosis (3)	Outliers (4)
1	2	0.2	-0.5	--
2	2	0.2	-0.7	--
3	2	-0.2	1.1	Top, Bottom
4	3	-0.2	-0.6	--
5	1	-2.2	5.8	Bottom
6	4	-2.2	5.8	Bottom
7	3	-0.1	-1.6	--
8	5	0.0	-1.2	--
9	2	-1.6	1.5	Bottom
10	3	-0.5	-1.0	--
11	3	0.3	-1.1	--
12	6	0.6	-1.1	---
13	3	-0.5	-0.9	--
14	3	1.6	6.3	Top, Bottom
15	3	1.8	2.4	Top
16	2	-0.2	-1.2	---
17	3	-1.2	1.5	Bottom
18	3	0.2	-0.8	---
19	4	0.7	-1.1	---
20	3	-1.1	1.1	Bottom
21	3	0.3	-1.9	--
22	3	-0.3	-0.3	--
23	3	-0.8	-1.0	--
24	2	-0.2	0.3	Bottom
25	1	0.5	0.1	--
26	1	-0.8	1.9	Bottom
Total	**73**	**-0.3**	**-0.8**	**--**

Kurtosis: Eleven out of 26 stations have positive kurtosis. The kurtoses of 3 stations (5, 6 & 14) were pronounced, highest in station 14 (6.3). Fifteen stations (1,2,4,7,8,10-13,16,18,19, 21-23 and the Internal Medicine total scores have negative kurtosis, lowest in station 21 at -1.9, (Test 3 in Table 4.20, Figure 4.5).

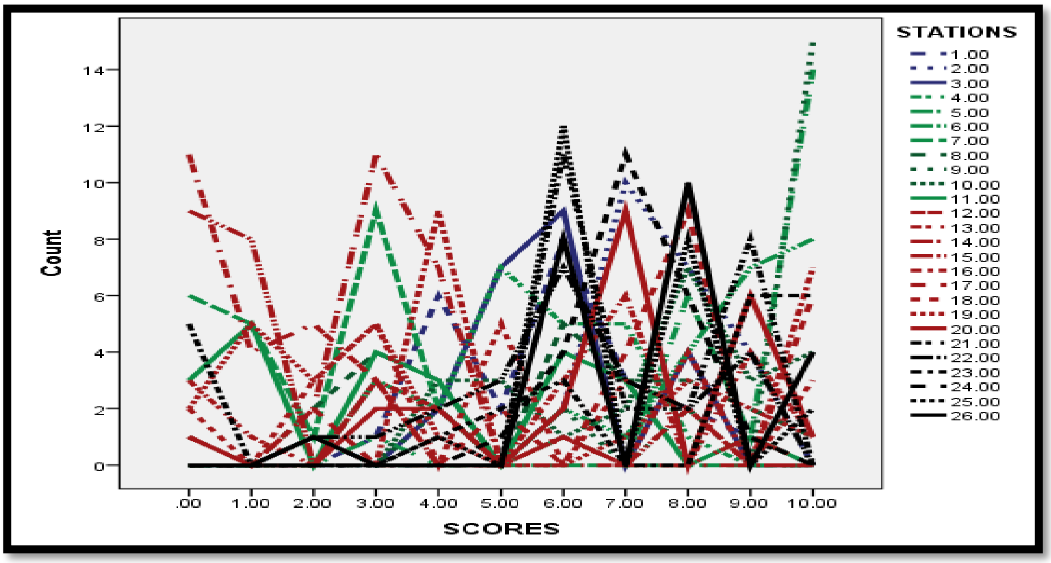

FIGURE 4.5: INTERNAL MEDICINE OSCE STATIONS SCORES DISTRIBUTION

Checking for outliers with Box-whiskers plot:

Ten stations had extreme scores. Out of these 10 stations, seven (5,6,9,17,20,24 & 26) have low extreme scores, while station 15 has large extreme scores. Stations 3 and 14 have both low and large extreme values (Test 4 in Table 4.20, Figure 4.6).

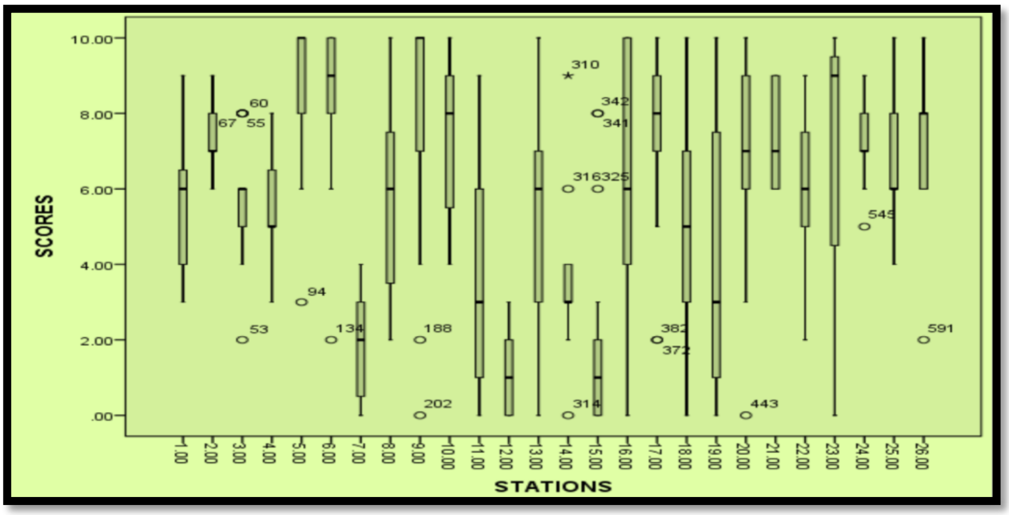

FIGURE 4.6: CHECKING FOR OUTLIERS IN INTERNAL MEDICINE OSCE STATIONS

Z-scores (cf. Table 4.21): After standardizing the raw scores, there were no extreme values. The Z-letter grades corresponded with different scores in each station. The best performance in Internal Medicine was good (grade C) and the lowest was in the very poor (grade F). The scores ranged from 4.7-6.5 in the total students' performance in Internal Medicine. The proportion of students in each grade of performance did not differ significantly from the normal distribution.

TABLE 4.21: SUMMARY OF INTERNAL MEDICINE Z-SCORES

GRADES	GOOD	AVERAGE	POOR	V.POOR	TOTAL
GRADES	C	D	E	F	
Z-SCORE RANGES	>+1<+2 (13.6%)	+1≤0≥-1 (68.3%)	<-1>-2 (13.6%)	-2≥-3 (2.2%)	100%
Frequency	6(26.1%)	14(60.9%)	2(8.7%)	1(4.3%)	23(100%)
Raw Score Equivalent	6.3-6.5	5.2-6.2	4.9-5.1	4.7	

4.7.1.2 *Measures of central tendencies*

All the measures of central tendencies (sum, mean, mode, median and standardised pass marks) in the internal medicine stations were highest in station 5 and lowest in station 12. The mean and median were equal in station 12. The standardised pass marks in station 3, 4 and 18 were the same as the university pass mark of 5, while 6 stations (7,11,12,14,15 & 19) pass marks were below 5. The remaining 17 stations have pass marks above 5, especially in stations 5 and 9 (cf. Table 4.22).

TABLE 4.22: INTERNAL MEDICINE OSCE STATIONS: MEASURES OF CENTRAL TENDENCIES

TESTS STATIONS	SUM (5)	MEAN (6)	MEDIAN (7)	MODE (8)	S. SETTING (9)
1	131	5.7	6	6	6
2	174	7.6	7	7	7.2
3	131	5.7	6	6	5
4	125	5.4	5	5	5
5	206	9.0	10	10	10
6	196	8.5	9	10	9
7	42	1.8	2	3	2
8	131	5.7	6	6	5.9
9	188	8.2	10	10	10
10	170	7.4	8	8	8
11	83	3.6	3	1	3
12	23	1.0	1	0	1
13	124	5.4	6	7	6
14	80	3.5	3	3	3
15	39	1.7	1	0	1
16	146	6.4	6	4	6
17	174	7.6	8	8	8
18	113	4.9	5	3	5
19	89	3.9	3	1	2.7
20	153	6.7	7	7	7
21	168	7.3	7	6	7
22	143	6.2	6	6	6.1
23	146	6.4	9	9	8
24	163	7.1	7	7	7
25	160	7.0	6	6	6.6
26	170	7.4	8	8	7.7
Total	**132.7**	**5.8**	**5.7**	**4.7**	**5.8**

4.7.1.3 *Measures of variability*

The range, standard deviation and standard error of the mean were widest in station 23 and narrowest in station 2 and the total scores (cf. Tests 10,11,12 & 14 in Table 4.23). The station tasks were significantly heterogeneous (cf. Test 16 in Table 4.23). The 95% confidence interval of the mean was found in 8 stations (1,4,8,13,16,18,19 & 23). The rest of the stations did not contain the university pass mark of 5 (cf. Test 13 in Table 4.23). The coefficient of variation was highest in station 15 and lowest in station 2 and the total scores.

TABLE 4.23: INTERNAL MEDICINE OSCE STATIONS: MEASURES OF VARIABILITY

TESTS STATIONS	S. DEVIATION (10)	RANGE (11)	SEM (12)	95% CI (13)	C. VARIATION (%) (14)
1	1.6	6	0.3	5-6.4	27.2
2	0.9	3	0.2	7.2-8	11.9
3	1.4	6	0.3	5.1-6.4	25.1
4	1.4	5	0.3	4.8-6	26
5	1.7	7	0.4	8.2-10	19.1
6	1.9	8	0.4	7.7-9.3	22.1
7	1.4	4	0.3	1.2-2.4	78.7
8	2.4	8	0.5	4.7-6.7	42.1
9	3.0	10	0.6	6.9-9.5	36.7
10	2.0	6	0.4	6.5-8.3	27.6
11	2.7	9	0.6	2.4-4.8	75.1
12	1.1	3	0.2	0.5-1.5	113
13	3.1	10	0.7	4-6.8	58.3
14	1.6	9	0.3	2.8-4.2	46.6
15	2.5	8	0.5	0.6-2.8	144.7
16	3.1	10	0.7	5-7.7	49
17	2.3	8	0.5	6.6-8.5	30
18	3.0	10	0.6	3.6-6.2	61.7
19	3.3	10	0.7	2.5-5.3	84.8
20	2.4	10	0.5	5.6-7.7	36.7
21	1.4	3	0.3	6.7-7.9	19.2
22	1.9	7	0.4	5.4-7.1	31.1
23	3.9	10	0.8	4.7-8	61.3
24	0.9	4	0.2	6.7-7.5	12.7
25	1.5	6	0.3	6.3-7.6	20.9
26	1.9	8	0.4	6.6-8.2	25
Total	**0.5**	**1.8**	**0.1**	**5.5-6**	**9.2**

ANOVA (comparing means): The variance between the stations was high (>30%). Moreover, the variance between the station scores was significantly higher than the variance within the stations. The mean of station 5 was significantly higher than the rest. (cf. Test 25 in Table 4.24).

TABLE 4.24: ANOVA TABLE FOR INTERNAL MEDICINE OSCE

VARIANCE	SS	*df*	MS	F	Sig
Between Groups	2570.9(47.4%)	25	102.8	20.6	0.00
Within Groups	2857(52.6%)	572	5		
Total	**5427.9(100%)**	**597**			
	Turkey's HSD Post Hoc test= All the Station means were significantly different from each other. Stations 5 and 12 means were significantly higher and lower than the rest of the stations respectively.				

SS*: Sum of squares; Df**:Degrees of freedom; MS***: Mean of Squares; F****:Ratio of between mean of squares and within mean of squares; SS*****:Significance

<u>Generalizability studies: variance components estimates</u>

The students and the examiners contributed 1.9% and 79.8% respectively of the variance obtained in the internal medicine test. The interaction between students and examiners was 18.4%. G-coefficient=0.1 (cf. Test 26 in Table 4.25).

TABLE 4.25: GENERALIZABILITY STUDIES FOR INTERNAL MEDICINE OSCE

	COMPONENTS	%			
Students	0.1	1.9%			
Examiners	4.3	79.8%			
Students*Examiners	4. 9/5examiners(1.0)	18.4%			
Total	**5.4**				
	G-coefficient= 0.1/1.1=0.1. Errors from examiners was 79.8%				

4.7.1.4 *Station analysis*

<u>T-test</u>: The standardised pass marks in stations 3, 4 and 18 were not different from the university pass mark of 5, because *t* was 0. The standardised pass marks in 6 stations (7, 11,12,14,15 & 19) were significantly below the university pass mark of 5 because *t* was negative. The standardised pass marks in the remaining 17 stations were significantly above the university pass mark of 5, because it was positive (cf. Test 15 in Table 4.26).

<u>Item Difficulty Index (IDI)</u>: Good IDI is between 0.3 and 0.8, furthermore, for a criterion referenced test, IDI of 0.9 is acceptable (cf. section 2.3.4.2.i.). Fifteen stations had IDI within the good range. Eight stations had IDI above the good range, while the remaining three stations had IDI below good range (cf. Test 17 in Table 4.26).

<u>Station Discrimination Index (*d*)</u>: The range of *d* is -1.00 to +1.00 (good range is between 0.3-0.5). Seven stations (11,14,16,19,20,22 & 23) had *d* within the good range (cf. Test 18 in Table 4.26). *d* in 10 stations (2,3,5,6,7,12,13,17,21 & 24) was 0. Two stations (8 & 18)

have *d* above the normal range. The rest of the stations have *d* below the normal range.

Statistical Significance: The mean of those who passed was lower than the mean of those who failed the Internal Medicine OSCE in 6 stations (3,5,13,14,17 & 25; cf. Test 19 in Table 4.26) and significantly so in stations 3 and 13. Pass means were not significantly higher than the Fail means in the other stations.

Failure rates**:** All the students failed in stations 7 and 12. No one failed in three stations (2,21 & 24; cf. Test 24 in Table 4.26).

TABLE 4.26: INTERNAL MEDICINE OSCE STATIONS ANALYSIS

TESTS STATIONS	T-test (Pmk) (15)	IDI (17)	*d* (18)	S. Sig (P/F) (19)	FAILURE RATES (%) (24)
1	9.1*	0.7	0.2	0.6, *p=0.5*	7(30.4)
2	20*	1	0	3.1*	0
3	0	0.9	0	-4.7*	3(13)
4	0	0.8	0.2	5.3*	5(21.7)
5	45.5*	1.0	0	-1.5, p=*0.1*	1(4.3)
6	36.4*	1.0	0	0.1, p=1.0	1(4.3)
7	-27.3*	0	0	6.7*	23(100)
8	7.8*	0.6	0.7	7*	9(39.1)
9	45.5*	0.8	0.2	2.0, p=0.1	4(17.4)
10	27.3*	0.9	0.2	3.5*	3(13)
11	-18.2*	0.4	0.3	5.9*	15(65.2)
12	-36.4*	0.0	0.0	2.2*	23(100)
13	9.1*	0.6	0.0	-2.5*	9(39.1)
14	-18.2*	0.1	0.3	-0.6, p=1.0	21(91.3)
15	-36.4*	0.1	0.2	3.4*	20(87)
16	9.1*	0.5	0.3	3.7*	11(47.8)
17	27.3*	0.9	0.0	-2.0, p=0.1	2(8.7)
18	0	0.6	0.7	3.1*	10(43.5)
19	-20.8*	0.3	0.3	3.6*	16(69.6)
20	18.2*	0.8	0.5	0.3, p=0.8	5(21.7)
21	18.2*	1.0	0.0	4.7*	0
22	10.4*	0.8	0.3	11.7*	4(17.4)
23	27.3*	0.7	0.5	1.8, p=0.1	6(26.1)
24	18.2*	1.0	0.0	6.7*	0
25	14.3*	1.0	0.2	-0.2, p=0.9	1(4.3)
26	24.6*	1.0	0.2	5.0*	1(4.3)
Total	**7***	**0.9**			**2(8.7)**

*P<0.05, significant

4.7.1.5 *Reliability checks*

Alpha-correlation with overall: The total alpha coefficient for the internal medicine stations was 0.4, which was low. Seven stations (3,5,6,9,16,17 & 23) had negative correlations with the total scores. The rest of the stations had positive, but low correlations with the total scores, the highest being 0.5 in station 8. Station 3 had the poorest correlation with total scores with -0.2 (cf. Test 20 in Table 4.27).

Alpha coefficient (if-item-deleted): When each station was deleted in turn, alpha in 23 stations improved, 9 of which (3,5,6,9,16,17,19,20 & 23) improved to or above the total alpha, while the alpha in stations 8, 10 and 22 decreased. (cf. Test 21 in Table 4.27).

Pearson's Correlation, r (station with student total) and r^2: Pearson's correlation *(r)* and contribution (r^2) to the total score variance was low in all the stations, the highest being 0.6 and 0.4 respectively in station 8. Stations 3, 5 and 17 had negative *r and r^2* (cf. Tests 22 & 23 in Table 4.27).

TABLE 4.27: INTERNAL MEDICINE OSCE STATIONS: RELIABILITY CHECKS

TESTS STATIONS	α CORRELATION (20)	α DELETED (21)	PEARSON CORR (r) (22)	r^2 (23)
1	0.2	0.3	0.3, $p=0.2$	0.1
2	0.3	0.3	0.3, $p= 0.1$	0.1
3	-0.2	0.4	-0.1, $p=0.6$	0.0
4	0.2	0.3	0.3, $p= 0.2$	0.1
5	-0.1	0.4	-0.0, $p= 1$	0.0
6	-0.1	0.4	0.1, $p=0.8$	0.0
7	0.3	0.3	0.4, $p=0.1$	0.2
8	0.5	0.2	0.6*	0.4
9	-0.1	0.4	0.1, $p=0.5$	0.0
10	0.3	0.3	0.5*	0.2
11	0.2	0.3	0.4, $p=0.1$	0.1
12	0.1	0.4	0.2, $p=0.5$	0.0
13	0.1	0.3	0.3, $p=0.1$	0.1
14	0.2	0.3	0.3, $p=0.2$	0.1
15	0.1	0.4	0.2, $p=0.3$	0.0
16	-0.1	0.4	0.1, $p= 0. 6$	0.0
17	-0.2	0.4	-0.1, $p=0.3$	0.0
18	0.2	0.3	0.4, $p=0.1$	0.2
19	0.1	0.4	0.3, $p=0.2$	0.1
20	0.0	0.4	0.2, $p=0.4$	0.0
21	0.2	0.3	0.3, $p=0.2$	0.1
22	0.4	0.3	0.5*	0.2
23	0.0	0.4	0.3, $p=0.2$	0.1
24	0.2	0.3	0.3, $p=0.2$	0.1
25	0.1	0.4	0.2, $p=0.5$	0.0
26	0.1	0.3	0.3, $p=0.2$	0.1
Total	**0.4**		**1**	**1.00**

$^*P<0.05$, significant

Guidelines for the interpretation of correlation coefficients: 0.75-1.00: strong; 0.50-0.74: moderate to high; 0.25-0.49: low to moderate; 0.00-0.24: weak.

4.7.1.6 *Identifying Hawks and Doves in Internal Medicine OSCE stations*

Station 5 was marked too leniently based on the following evidences:

- The scores distribution is negatively skewed (cf. Test 2 in Table 4.20, Figure 4.5)
- Had excess kurtosis (cf. Test 3 in Table 4.20, Figure 4.5).
- Had outliers (cf. Test 4 in Table 4.20, Figure 4.5).
- Had significantly higher mean (cf. Test 25 in Table 4.24).
- Variability was small (cf. Tests 10-16 in Table 4.23).
- IDI was too high (cf. Test 17 in Table 4.26).
- Discriminating power was very poor (cf. Test 18 in Table 4.26).
- Correlations were negative and contribution was low to the total student performance in the OSCE (cf. Test 22-23 in Table 4.27).

4.8 REPORT ON STATION METRICS IN SURGERY

In the Surgery OSCE, no student attempted any of the tasks in station 23. The students claimed they were not familiar with the topic tested. However, the facilitator of the subject informed the faculty that the students were taught the topic, but the records showed that very few students attended the lecture. The report on station metrics in surgery is as follows.

4.8.1 Surgery stations metrics

4.8.1.1 *Frequency distribution*

Skewness: All the stations' score distribution was skewed (cf. Test 2 in Table 4.28, Figure 4.7). Eighteen stations and the total scores (1-4,6,8,10-15,17-20,24 & 25) were negatively skewed. The rest of the stations had positive skewness, highest was 0.7 in station 22. The lowest negative skewness was in station 1 with -1.0.

TABLE 4.28: SURGERY OSCE STATIONS SCORES DISTRIBUTION

TESTS STATIONS	NUMBER OF TASKS (1)	SKEWNESS (2)	KURTOSIS (3)	OUTLIERS (4)
1	3	-1.0	0.7	Bottom
2	2	-0.7	-0.2	Bottom
3	2	-0.1	-0.7	--
4	1	-0.9	0.0	--
5	2	0.1	-0.6	--
6	3	-1.0	0.7	Bottom
7	1	0.2	-0.8	--
8	1	-0.5	-0.1	--
9	1	0.2	-0.2	--
10	1	-0.7	0.3	Bottom
11	1	-0.7	-0.1	Bottom
12	1	-1.1	0.5	Bottom
13	1	-0.5	-0.3	--
14	1	-0.6	0.2	--
15	1	-0.2	0.7	Top, Bottom
16	1	0.2	-0.6	--
17	1	-0.3	0.2	--
18	1	-0.3	-1.1	--
19	1	-0.7	-0.4	Bottom
20	1	-0.2	-1.1	--
21	1	0.1	-1.7	--
22	1	0.7	-0.9	--
23	2	0.0	0.0	--
24	1	0.0	-0.9	---
25	1	-0.7	1.5	Top, Bottom
Total	**33**	**-0.4**	**-0.6**	**--**

Kurtosis: Ten stations had positive kurtosis, the highest was 1.5 in station 25. Fifteen stations and the surgery total scores had negative kurtosis (2,3,5,7-9,11,13,16,18-22 & 24), lowest was in station 21 at -1.7, (Test 3 in Table 4.28, Figure 4.7).

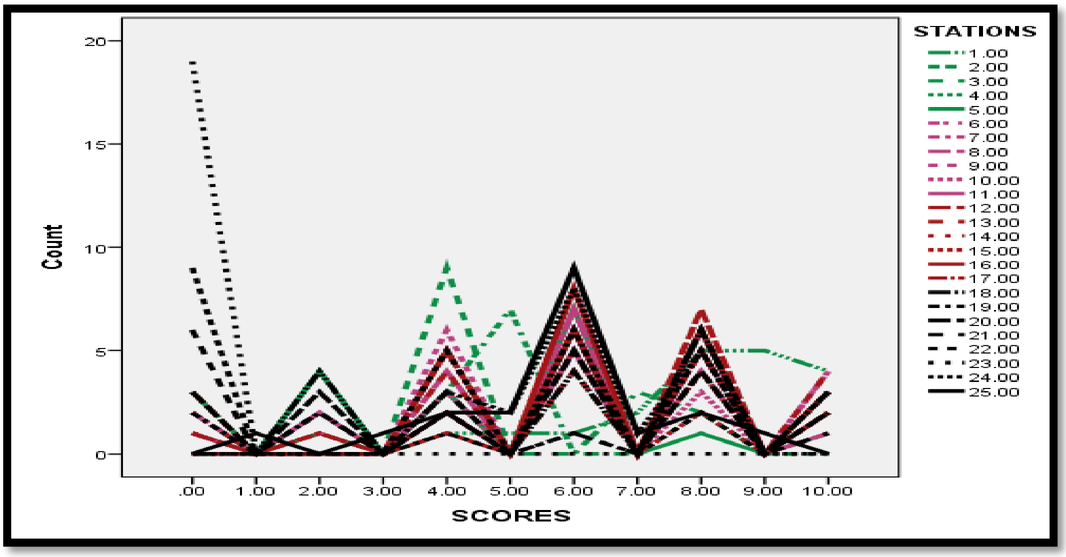

FIGURE 4.7: SURGERY OSCE STATIONS SCORES DISTRIBUTION

Checking for outliers with Box-whiskers plot

Seven stations (1,2,6,10-12 & 19) had low extreme scores (Test 4 in Table 4.28, Figure 4.8). Two stations (15 and 25) had both large and low extreme scores. The rest of the stations had no extreme values.

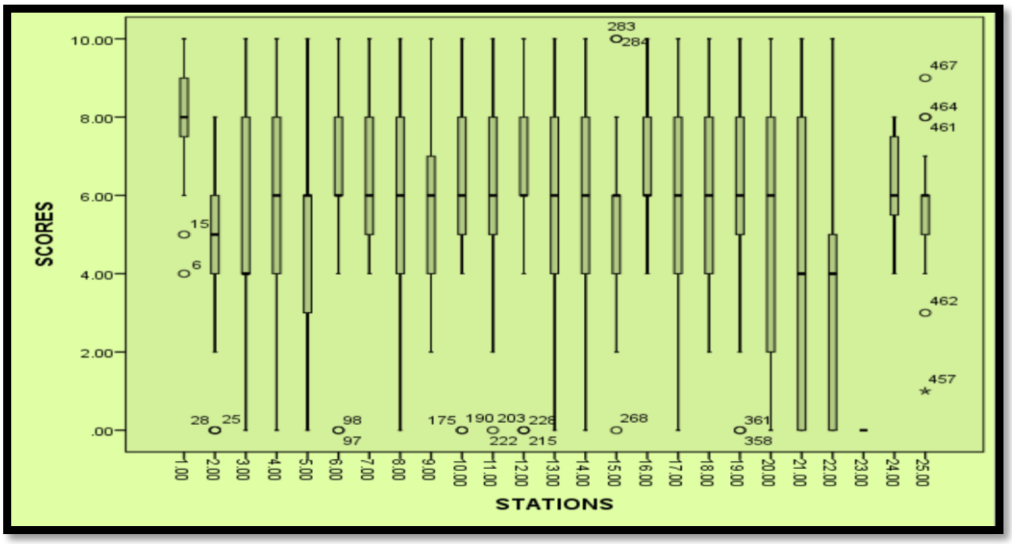

FIGURE 4.8: CHECKING FOR OUTLIERS IN SURGERY STATIONS

Z-scores (cf. Table 4.29): After standardizing the raw scores, there were no extreme values. The Z-letter grades corresponded with different scores in each station. The best performance in surgery was good (grade C) and the worst was poor (grade E) with the scores ranging from 3.7-7.2.

TABLE 4.29: SUMMARY OF SURGERY Z-SCORES

GRADES	GOOD	AVERAGE	POOR	TOTAL
Grades	C	D	E	
Z-score ranges	>+1<+2 (13.6%)	+1≤0≥-1 (68.3%)	<-1>-2 (13.6%)	100%
Frequency	2(10.5%)	14(73.7%)	3(15.8%)	19(100%)
Raw Score Equivalent	6.7-7.2	4.7-6.5	3.7-4	

4.8.1.2 *Measures of central tendencies in Surgery stations*

All the measures of central tendencies (sum, mean, mode, median and standardised pass marks) in the surgery stations were highest in station 1 and lowest in station 22, (cf. Tests 5,6,7,8 & 9 in Table 4.30). The mean, median and mode were equal only in station 19. However, these measures were in proximity to each other in other stations except in stations 18, 21 and 22. The standardised pass marks in station 2 was the same as the university pass mark of 5, while in 20 stations (1,4-20,24-25) and the total scores, the pass marks were above 5, especially in station 1. The pass marks in the remaining 4 stations were below 5, especially in station 22.

TABLE 4.30: SURGERY OSCE STATIONS: MEASURES OF CENTRAL TENDENCIES

TESTS STATIONS	SUM(5)	MEAN (6)	MEDIAN (7)	MODE (8)	S. SETTING (9)
1	154	8.1	8	8	8.4
2	86	4.5	5	5	5
3	96	5.1	4	4	4.6
4	106	5.6	6	6	6.3
5	100	5.3	6	6	5.4
6	124	6.5	6	6	6.9
7	120	6.3	6	6	6.3
8	106	5.6	6	6	5.7
9	110	5.8	6	6	5.7
10	118	6.2	6	6	6.3
11	120	6.3	6	8	6.9
12	116	6.1	6	8	6.9
13	122	6.4	6	8	6.9
14	110	5.8	6	4	6
15	106	5.6	6	6	5.7
16	134	7.1	6	6	6.9
17	110	5.8	6	6	5.7
18	118	6.2	6	8	6.9
19	114	6	6	6	6.6
20	96	5.1	6	6	5.4
21	80	4.2	4	0	4.3
22	62	3.3	4	0	2.3
23	0	0	0	0	0
24	117	6.2	6	6	6
25	108	5.7	6	6	6
Total	**105**	**5.6**	**5.7**	**3.7**	**5.7**

4.8.1.3 *Measures of variability*

The range, standard deviation, standard error of the mean and coefficient of variation were widest in station 22 and narrowest in station 24 (cf. Tests 10,11,12 & 14 in Table 4.31). The station tasks were significantly heterogeneous (Levene statistics=3.6, p<0.05). The 95% confidence interval of the mean in all the stations contain the university pass mark of 5, except in 3 stations (1,16 & 24) and the total scores (cf. Test 13 in Table 4.31).

TABLE 4.31: SURGERY OSCE STATIONS: MEASURES OF VARIABILITY

TESTS STATIONS	S. DEVIATION (10)	RANGE (11)	SEM (12)	95% CI (13)	C. VARIATION (%) (14)
1	1.7	6	0.4	7.3-8.9	21
2	2.5	8	0.6	3.3-5.7	55.2
3	2.9	10	0.7	3.7-6.4	56.9
4	3.0	10	0.7	4.2-7	52.7
5	2.9	10	0.7	3.9-6.7	55.1
6	3.0	10	0.7	5-8	45.5
7	1.8	6	0.4	5.4-7.2	28.5
8	2.6	10	0.6	4.4-6.8	45.7
9	2.0	8	0.5	4.8-6.7	34.5
10	3.0	10	0.7	4.8-7.6	47.9
11	2.9	10	0.7	4.9-7.7	45.1
12	3.1	10	0.7	4.6-7.6	50.6
13	2.9	10	0.7	5-7.8	44.7
14	2.8	10	0.7	4.4-7.1	48.7
15	2.5	10	0.6	4.4-6.8	44.1
16	1.8	6	0.4	6.2-7.9	25.7
17	2.6	10	0.6	4.5-7	44.4
18	2.8	8	0.7	4.9-7.6	45.4
19	3.1	10	0.7	4.5-7.5	52.2
20	3.3	10	0.8	3.5-6.6	64.5
21	3.7	10	0.8	2.5-6	86.7
22	3.7	10	0.8	1.5-5	112.3
23	0.0	0	0.0	0	0
24	1.4	4	0.3	5.5-6.8	22.4
25	1.8	8	0.4	4.8-6.6	32.2
Total	**1.0**	**3.54**	**0.2**	**5.1-6**	**18.2**

ANOVA (comparing means): The variance between the stations was less than 30%. However, the variance between the station scores was significantly higher than the variance within the stations. The mean of station 3 was significantly higher than the rest. (cf. Table 4.32).

TABLE 4.32: ANOVA TABLE FOR SURGERY

VARIANCE	SS	*Df*	MS	F	Sig
Between Groups	1008.3(24.1%)	24	42	5.96	0.00
Within Groups	3171.6(75.9%)	450	7		
Total	**4179.9(100%)**	**474**			
	Turkey's HSD Post Hoc test= Many of the Station means were not significantly different from each other. Stations 1 and 22 means were significantly higher and lower than the mean of some of the stations respectively. Station 23 has 0 mean.				

SS*: Sum of squares; Df**:Degrees of freedom; MS***: Mean of Squares; F****:Ratio of between mean of squares and within mean of squares; SS*****:Significance

Generalizability studies: variance components estimates

The students and the examiners contributed 21.3% and 50.7% respectively of the variance obtained in the surgery test. The interaction between students and examiners was 28%. G-coefficient=0.43 (cf. Table 4.33).

TABLE 4.33: GENERALIZABILITY STUDIES FOR SURGERY OSCE

	COMPONENTS	%
Students	0.8	21.3%
Examiners	1.9	50.7%
Students*Examiners	6.26/6 Examiners(1.0)	28%
Total	**3.7**	
	G-coefficient= 0.79/1.84=0.43. Errors from examiners was 50.7%	

4.8.1.4 *Station analysis of the Surgery OSCE stations*

T-test: The standardised pass marks in four stations (2,3,5 & 20) were not significantly different from the university pass mark of 5 (cf. Test 15 in Table 4.34). The rest of the stations had significantly different pass marks from the university pass mark of 5.

Item Difficulty Index (IDI): Good IDI is between 0.3 and 0.8, furthermore, for a criterion referenced test, IDI of 0.9 is acceptable (cf. Chapter 2, section 2.5.4.4). Twenty-two stations (2-15,16-21 & 24-25) and the total surgery scores had IDI within the good range. Station 1 IDI was above, while station 22 IDI was below the good range respectively (cf. Test 17 in Table 4.34).

Station Discrimination Index (*d*): The range of *d* is -1.00 to +1.00 (good range is between 0.3-0.5). Eight stations (2,5,8,10,11,13,18 & 20) had *d* within the good range (cf. Test 18 in Table 4.34). *d* in stations 7 and 22 were negative. Stations 4 and 25 had 0 *d,* while *d* was 1 in station 14 and the total surgery scores.

Statistical Significance: The mean of those who passed were significantly different from the mean of those who failed the surgery OSCE in all the stations except in station 23, where the students did not attempt any of the tasks (cf. Test 19 in Table 4.34).

Failure rates: The highest failure rate was found in station 22 with 14 (73.7%) and the lowest in station 1 (cf. Test 24 in Table 4.34).

TABLE 4.34: SURGERY OSCE STATIONS ANALYSIS

TESTS STATIONS	T-TEST (Pmk) (15)	IDI (17)	*d* (18)	S. SIG (P/F) (19)	FAILURE RATES (%) (24)
1	14.9*	1.0	0.2	13*	1(5.3)
2	0.0	0.6	0.4	10.8*	7(36.8)
3	-1.9	0.4	0.2	17.9*	12(63)
4	5.6*	0.7	0	14.5*	6(31.5)
5	1.9	0.6	0.4	8.6*	8(42)
6	8.1*	0.8	0.2	13.3*	4(21)
7	5.6*	0.7	-0.2	9.1*	5(26.3)
8	3.1*	0.6	0.4	11.2*	7(36.8)
9	3.1*	0.6	0.6	8.5*	7(36.8)
10	5.6*	0.7	0.4	10.8*	5(26.3)
11	8.1*	0.7	0.4	12.9*	5(26.3)
12	8.1*	0.8	0.6	17.8*	4(21)
13	8.1*	0.7	0.4	11.0*	6(31.6)
14	4.4*	0.6	1	10.7*	7(36.8
15	3.1*	0.6	0.6	8.4*	7(36.8
16	8.1*	0.9	0.2	9.1*	2(10.5)
17	3.1*	0.6	0.6	9.3*	7(36.8)
18	8.1*	0.7	0.4	12.3*	6(31.6)
19	6.8*	0.7	0.6	13.9*	5(26.3)
20	1.9	0.6	0.4	12.0*	8(42)
21	3.1*	0.4	0.2	17.3*	11(57.9)
22	11.8*	0.3	-0.2	9.3*	14(73.7)
23	-21.7*	0.0	0	0.0	19(100)
24	4.4*	0.8	0.2	9.4*	3(15.8)
25	4.4*	0.8	0	11.7*	4(21)
Total	**3.0***	**0.7**	**1**	**11***	**5(26.3)**

**P<0.05, significant*

4.8.1.5 *Reliability checks*

Alpha-correlation with overall: The total alpha coefficient for the surgery stations was 0.76, which was strong. Station 22 had negative correlation with the total surgery scores. Stations 12 and 14 had strong positive correlations with the total scores. The rest of the stations had low-moderate correlations with the total surgery scores (cf. Test 20 in Table 4.35).

Alpha coefficient (if-item-deleted): When each station was deleted in turn, alpha in 22 stations improved. The alpha-deleted improved to or above the total alpha in 9 stations (1-4,6-7,21-22 & 25), while the alpha in stations 12 and 14 decreased. (cf. Test 21 in Table 4.35).

Pearson's Correlation, *r* (station with student total) and r^2: Pearson's correlation and contribution to the total score variance was negative and very poor in stations 7 and 22.

The highest and significantly important correlations and contributions were in stations 12 and 14 (cf. Tests 22 & 23 in Table 4.35). Correlations and contributions were moderately-low in the other stations.

TABLE 4.35: SURGERY OSCE STATIONS: RELIABILITY CHECKS

TESTS STATIONS	α CORRELATION (20)	α DELETED (21)	PEARSON CORR (*r*) (22)	r^2 (23)
1	0.3	0.8	0.3, *p*=0.2	0.1
2	0.2	0.8	0.3, p=0.2	0.1
3	0.1	0.8	0.1 p=0.4	0.0
4	0.1	0.8	0.2, p=0.3	0.1
5	0.3	0.8	0.4, p=0.1	0.1
6	0.2	0.8	0.3, p=0.2	0.1
7	-0.2	0.8	-0.2, p=0.5	0.0
8	0.3	0.8	0.4, p=0.1	0.1
9	0.3	0.8	0.4, p=0.1	0.1
10	0.3	0.8	0.4, *p*=0.1	0.2
11	0.3	0.8	0.4, p=0.1	0.2
12	0.8	0.7	0.8*	0.7
13	0.6	0.7	0.7*	0.4
14	0.8	0.7	0.8*	0.6
15	0.5	0.7	0.5*	0.3
16	0.5	0.7	0.6*	0.3
17	0.5	0.7	0.5*	0.3
18	0.5	0.7	0.5*	0.3
19	0.6	0.7	0. 7*	0.5
20	0.5	0.7	0.6*	0.3
21	0.1	0.8	0.3, p=0.3	0.1
22	-0.3	0.8	-0.2, p=0.4	0.1
23	--	--	---	--
24	0.4	0.8	0.4, p=0.1	0.2
25	0.1	0.8	0.2, p=0.4	0.0
Total	**0.8**	**---**	**1.00**	**1**

**P<0.05, significant*

Guidelines for the interpretation of correlation coefficients: 0.75-1.00: strong; 0.50-0.74: moderate to high; 0.25-0.49: low to moderate; 0.00-0.24: weak.

4.8.1.6 *Identifying Hawks and Doves in Surgery OSCE stations*

Stations 1 and 22 were marked too leniently and too strictly respectively based on the following evidences:

- The scores distribution was negatively skewed in station 1 and positively skewed in station 22 (cf. Test 2 in Table 4.28, Figure 4.7).

- Kurtosis was positive in station 1 and negative in station 22 (cf. Test 3 in Table 4.28, Figure 4.7).
- Station 1 had low extreme value, while station 22 has none (cf. Test 4 in Table 4.28, Fig 4.8).
- Station 1 had significantly higher mean, while station 22 had significantly lower mean than the rest of the stations (cf. Test 25 in Table 4.32).
- Variability was small in station 1 but excess in station 22 (cf. Tests 10-16 in Table 4.31).
- IDI was too high in station 1 and low in station 22 (cf. Test 17 in Table 4.34).
- Discriminating power was very poor in station 1 and negative in station 22 (cf. Test 18 in Table 4.34).
- Alpha deleted is in excess of the total alpha in both stations (cf. Tests 22-23 in Table 4.35).
- Correlations and contributions were low in both stations (cf. Tests 22-23 in Table 4.35).

4.9 OVERALL REPORT ON STATION METRICS IN THE JULY 2015 OSCE

The overall report on station metrics in the July 2015 OSCE is as follows:

4.9.1 Overall stations metrics

4.9.1.1 *Frequency distribution*

Skewness: All the subjects' score distribution was skewed (cf. Test 2 in Table 4.36, Figure 4.9). The total internal medicine and surgery scores were negatively skewed. The lowest negative skewness was in surgery with -0.4, while the highest positive skewness was in OBGYN with 0.5.

TABLE 4.36: JULY 2015 OSCE STATIONS SCORES DISTRIBUTION

TESTS	OBGYN	PAED	INT MED	SURG	OVERALL
1. Number of Tasks	10	56	73	33	172
2. Skewness	0.5	0.1	-0.3	-0.4	0.2
3. Kurtosis	-0.8	-1.1	-0.8	-0.6	-1.3
4. Outliers	—	--	---	--	--

Kurtosis: All the subjects had negative kurtosis, lowest in internal medicine with -0.8. The kurtosis in the OSCE overall was -1.3, (Test 3 in Table 4.36, Figure 4.9).

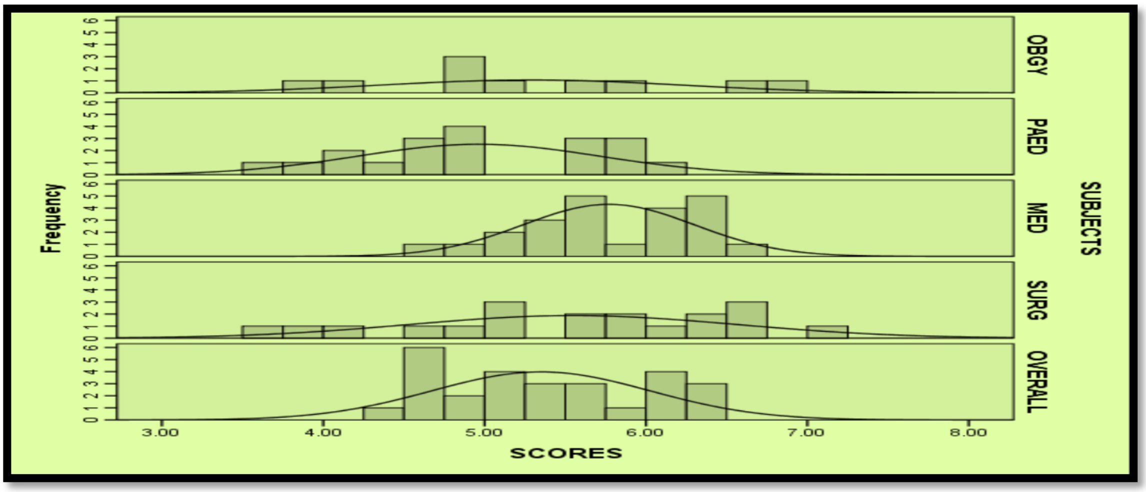

FIGURE 4.9: SCORES DISTRIBUTION IN THE JULY 2015 FINAL OSCE

<u>Checking for outliers with Box-whiskers plot</u>:

There were no outliers in the overall and subject OSCE scores (cf. Test 4 in Table 4.36, Figure 4.10).

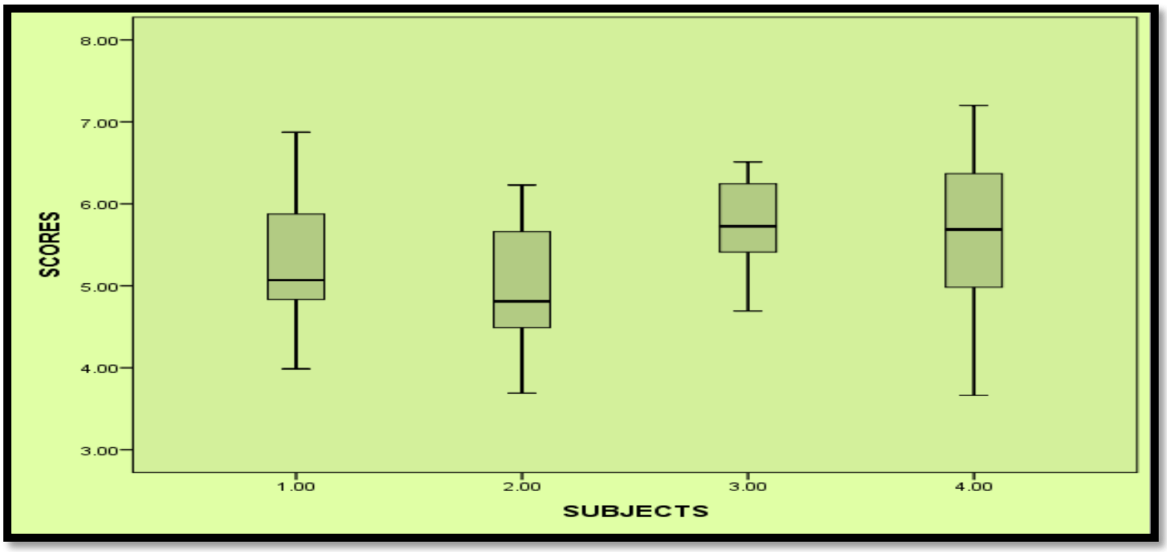

FIGURE 4.10 CHECKING FOR OUTLIERS IN THE JULY 2015 OSCE STATIONS (SUBJECTS: 1=OBGYN; 2=PAED; 3=INT MED; 4=SURGERY)

<u>Z-scores</u> (cf. Table 4.37): After standardizing the raw scores, there were no extreme values. The Z-letter grades corresponded with different scores in each subject. The best performance in the overall was good (grade C) and the worst was poor (grade E), with scores ranging from 4.4 to 6.4.

TABLE 4.37: SUMMARY OF Z-SCORES IN THE JULY 2015 OSCE

GRADES	GOOD	AVERAGE	POOR	TOTAL
GRADES	C	D	E	
Z-SCORE RANGES	>+1<+2 (13.6%)	+1≤0≥-1 (68.3%)	<-1>-2 (13.6%)	100%
Frequency	7(25.9%)	14(51.9%)	6(22.2%)	27(100%)
Raw Score Equivalent	6.1-6.4	4.7-5.8	4.4-4.6	

4.9.1.2 *Measures of central tendencies*

All the measures of central tendencies (sum, mean, mode, median and standardised pass marks) in the July 2015 OSCE were highest in internal medicine. Paediatric had the lowest mean, median and standard pass mark. OBGYN had the lowest mode of 3.7 (cf. Tests 5,6,7,8 & 9 in Table 4.38). The mean, median and mode were not equal in any of the subject-totals; however, these measures were in proximity to each other except in surgery. None of the subjects had a standard pass mark of 5 (the fixed university pass mark). Only paediatrics had a standard pass mark below 5. The pass marks in the other subjects were above 5, especially in Internal Medicine.

TABLE 4.38: JULY 2015 FINAL OSCE: MEASURES OF CENTRAL TENDENCIES

TESTS	OBGYN	PAED	INT MED	SURG	OVERALL
5. Sum	54.3	94.3	132.7	105	144.3
6. Mean	5.4	5.0	5.8	5.6	5.4
7. Median	5.3	4.8	5.7	5.7	5.3
8. Mode	4.3	3.7	4.7	3.7	4.5
9. S. Setting	5.3	5.0	5.8	5.7	5.3

4.9.1.3 *Measures of variability*

The range, standard deviation, standard error of the mean and coefficient of variation were widest in surgery, followed by OBGYN. Variation was narrowest in Internal Medicine (cf. Tests 10,11,12 & 14 in Table 4.39). The station tasks were significantly heterogeneous in all the subjects (cf. Test 16 in Table 4.39). The 95% confidence interval of the mean in internal medicine, surgery and the overall July OSCE scores did not contain the university pass mark of 5 (cf. Test 13 in Table 4.39).

TABLE 4.39: JULY 2015 FINAL OSCE: MEASURES OF VARIABILITY

TESTS	OBGYN	PAED	INT MED	SURG	OVERALL
10. S. deviation	1.0	0.8	0.5	1.0	0.7
11. Range	2.8	2.5	1.8	3.5	2
12. SEM	0.3	0.2	0.1	0.2	0.1
13. 95% CI(Mean)	4.8-6.1	4.6-5.3	5.5-6	5.1-6	5.1-5.6
14. C. Variation (%)	17.5	15.1	9.2	18.2	13
16. Homogeneity	L=2.4, *p*=0.1. The tasks in the different subjects are homogenous				

ANOVA (comparing means): The variance between the subject scores was low (<30%) but significantly higher than the variance within the subjects. The overall mean in Medicine was significantly higher than the mean in Paediatrics. The mean in Paediatrics was the lowest (cf. Test 25 in Table 4.40).

TABLE 4.40: OVERALL ANOVA TABLE FOR THE JULY 2015 FINAL OSCE

VARIANCE	SS	*df*	MS	F	Sig
Between Groups	7.3(14.6%)	3	2.4	3.8	0.0
Within Groups	42.7(85.4%)	67	0.6		
Total	**50(100%)**	**70**			
	Turkey's HSD Post Hoc test= The overall mean in medicine was significantly higher than the mean in Paediatrics				

SS*: Sum of squares; Df**: Degrees of freedom; MS***: Mean of Squares; F****:Ratio of between mean of squares and within mean of squares; SS*****:Significance

Generalizability studies: variance components estimates

The students and subjects contributed 54.3% and 25.4% respectively to the variance obtained in the July 2016 OSCE. The error from the interaction between the students and the tests was 20.3%. G-coefficient=0.75 (cf. Test 26 in Table 4.41).

TABLE 4.41: GENERALIZABILITY STUDIES: VARIANCE COMPONENTS ESTIMATES

	COMPONENTS	%
Students	0.3	54.3%
Subjects	0.1	25.4%
Students*Subject	0.4/4 subjects(0.1)	20.3%
Total	**0.5**	
	G-coefficient= 0.3/0.4=0.75. Errors from tests was 25.4%	

4.9.1.4 *Station analysis*

T-test: The standard pass marks in OBGYN and Paediatrics were not significantly different from the university pass mark of 5 (cf. Test 15 in Table 4.42). The pass marks for the other 2 subjects and the overall were significantly different from the university pass mark of 5.

Item Difficulty Index (IDI): Good IDI is between 0.3 and 0.8, furthermore, for a criterion referenced test, IDI of 0.9 is acceptable (cf. Section 2.5.4.4). Hence, all the subjects had IDI within the good range. (cf. Test 17 in Table 4.42).

Station Discrimination Index (*d*): The range of *d* is -1.0 to +1.0 (good range is between 0.3-0.5). OBGYN and paediatrics had *d* below the good range, while Internal Medicine and Surgery had *d* above the good range (cf. Test 18 in Table 4.42).

Statistical Significance: The means of those who passed were significantly higher than the means of those who failed the July 2016 OSCE in all the subjects except in Paediatrics (cf. Test 19 in Table 4.42).

Failure rates: The highest failure rate was found in paediatrics with 12 (63.2%) and lowest in Internal Medicine, 2 (8.7%) (cf. Test 24 in Table 4.42).

TABLE 4.42: JULY 2015 FINAL OSCE: STATIONS ANALYSIS

TESTS	OBGYN	PAED	INT MED	SURG	OVERALL
15. T-test (Pmk)	t=0.9,*p*>0.0	-0.5, *p*>0.0	7*	3.0*	2.3*
17. IDI	0.7	0.4	0.9	0.7	0.7
18. *d*	0.3	0.3	0.6	0.9	--
19. S. Sig (P/F)	4.6*	1.0, *p*=0.3	9.2*	7.3*	--
24. Failure rates (%)	3(30%)	12(63.2)	2(8.7)	5(26.3)	9(33)

**P<0.05, significant*

4.9.1.5 *Reliability checks*

Alpha-correlation with overall**:** The overall alpha coefficient for the July 2015 final OSCE was 0.2, which was very poor. Alpha correlations with the overall scores were positive but low in all the subjects especially in OBGYN (cf. Test 20 in Table 4.43).

Alpha coefficient (if-item-deleted): When each subject was deleted in turn, alpha in internal medicine and surgery improved and that of OBGYN improved to above the overall alpha, while the alpha in paediatrics decreased (cf. Test 21 in Table 4.43).

Pearson's Correlation, *r* (station with student total) and r^2**:** Pearson's correlations and contributions to the overall score variance were low in all the subjects especially in Surgery (cf. Tests 22 & 23 in Table 4.43).

TABLE 4.43: JULY 2015 FINAL OSCE: RELIABILITY CHECKS

TESTS	OBGYN	PAED	INT MED	SURG	OVERALL
20. α Correlation	0.1	0.2	0.2	0.1	0.2
21. α Deleted	0.3	0.1	0.2	0.2	---
22. Pearson Corr (*r*)	0.2, *p*=0.6	0.3, *p*=0.2	0.3, *p*=0.2	0.2, *p*=0.6	1
23. r^2	0.0	0.1	0.1	0.0	1

Guidelines for the Interpretation of correlation coefficients: 0.75-1.00: strong; 0.50-0.74: moderate to high; 0.25-0.49: low to moderate; 0.00-0.24: weak.

4.9.1.6 *Identifying Hawks and Doves in the overall OSCE scores*

Internal medicine scoring was Dove-like based on the following evidences:

- The scores distribution were skewed (cf. Test 2 in Table 4.36, Figure 4.9)
- Had significantly higher mean (cf. Test 25 in Table 4.40).
- Variability was small (cf. Tests 10-16 in Table 4.39).
- IDI was too high (cf. Test 17 in Table 4.42).
- Discriminating power was very poor (cf. Test 18 in Table 4.42).
- Correlations and contribution were low (cf. Tests 22-23 in Table 4.43).

4.10 REPORT ON OTHER ATTRIBUTES OF THE JULY 2015 OSCE

This Section briefly investigated the relationship, if any between the scores of the manned and written stations, OSCE psychometric properties and factors such as the candidate gender and the level of examiners that participated in the OSCE. The quality properties of the borderline test that was used to determine the OSCE station pass marks in this study was also described.

4.10.1 *Correlation(r) between the manned and written stations' scores*

Comparison was made between the mean scores of the manned and written stations for each subject using the Pearson's correlation (*r*). The correlations between the manned and written stations' scores were very strong in all the subjects except OBGYN. In OBGYN, the correlation was 0 (in other words, there was no correlation) between the manned and written stations. Applying the t-test on the relationship between the OBGYN manned and written mean scores, it was discovered that the mean score of the manned stations was significantly higher than the mean score of the written stations (t=-17.4, $p<0.05$) (cf. Table 4.44).

TABLE 4.44: CORRELATION(*r*) BETWEEN THE MANNED AND WRITTEN STATIONS' SCORES

TESTS	OBGYN	PAED	INT MED	SURG
Written(mean)	4.0	3.3	4.2	3.9
Manned(mean)	6.6	3.0	5.3	2.6
Coefficient of variation	0.1(W),0.3(M)	0.8(W),0.8(M)	0.6(W),0.6(M)	0.8(W),0.9(M)
r value	0.0	0.9	1.0	0.9
Significance	1.0	0.0	0.0	0.0

4.10.2 *Comparing candidate gender and means per subject*

In Table 4.45 below, the mean of the male candidates were significantly lower than the mean of the female candidates in paediatrics only. Paediatrics had 2 male and 2 female examiners. In the other departments, the gender mean was not significantly different.

TABLE 4.45: COMPARING STUDENT GENDER AND MEANS PER SUBJECT

TESTS	OBGYN	PAED	INT MED	SURG
Male student Mean	5.3	4.9	5.7	5.6
Female student Mean	5.3	5.5	5.9	5.3
T value	-0.2	-3.5	-1.7	1.1
Significance	0.9	0.0	0.1	0. 3

4.10.3 Comparing the academic level of the examiners and the G-coefficient

In Table 4.46 below, the examiners' error was highest and the G-coefficient was lowest in OBGYN. However, the interaction between the students and the examiners was strongest in the surgery, where the number of junior assessors (medical officers) was highest. The correlation between the number of medical officers and the number of examiners' errors was negative. There was no significant relationship between the number of medical officers who participated in the OSCE and the level of examiners errors, interactions between students and examiners and G-coefficient.

TABLE 4.46: COMPARING THE ACADEMIC LEVEL OF THE EXAMINERS AND THE G-COEFFICIENT
(Table continues on next page)

DEPARTMENT	NUMBER OF JUNIOR LEVEL EXAMINERS	EXAMINERS' ERROR	INTERACTION BETWEEN STUDENTS AND EXAMINERS	G-COEFFICIENT
OBGYN	3	89.2%	10.7%	0.0
Paediatrics	2	65.8%	26.5%	0.22
Internal Medicine	2	79.8%	18.4%	0.1
Surgery	5	50.7%	28%	0.43

Pearson's(*r*)		-0.6	0.3	0.8
Significance		0.4	0.7	0.2
Importance(r^2)		0.4	0.1	0.7

4.10.4 Evaluation of the borderline method of standard setting

The borderline method was used to determine the pass marks in this OSCE. In Table 4.47 below, the sensitivity of the test was moderate, while the specificity was very high. The predictive value for passing was high and that for failing was moderately high.

TABLE 4.47: EVALUATION OF THE BORDERLINE METHOD OF STANDARD SETTING

ITEM	UNIVERSITY PASS	UNIVERSITY FAIL	TOTAL
Standard Setting Pass	31	2	33
Standard Setting Fail	16	22	38
Total	47	24	71

- Sensitivity of borderline method of Standard Setting=31/47=0.7.
- Predictive value for passing=31/33=0.9
- Specificity of borderline method of Standard Setting=22/24=0.9.
- Predictive value for failing=22/38=0.6

4.11 CONCLUSION

Chapter 4 provided a report of the psychometric analysis of the OSCE's including its interactions with candidate gender and the academic level of the examiners in this study. The quality properties of the borderline methods of determining the pass marks of the OSCE stations involved in this study was also described. In the next chapter, Chapter 5, **Interpretation and discussion of the report on the Psychometric analysis**, other properties of the OSCE's and the Borderline Method used in this study is presented.

CHAPTER 5

INTERPRETATION AND DISCUSSION: PSYCHOMETRIC ANALYSIS OF THE OSCEs

5.1 INTRODUCTION

The purpose of this chapter is to interpret and discuss the findings in the psychometric analysis of the final OSCEs conducted in July 2015 at Kampala International University (KIU) Dar es Salaam campus, an example of a resource limited institution in East Africa. Recall that the OSCEs were done in four clinical departments (OBGYN, Paediatrics, Internal Medicine and Surgery) and was reported in Chapter 4. The analysis is organised as follows:

5.2 THE PSYCHOMETRIC TESTS REPORT

The psychometric tests were in form of descriptive and inferential statistics. The descriptive statistics included distribution of the scores, measures of centre and standard setting as well as measures of variability. Whereas the inferential statistics consist of station analysis, reliability estimates, identifying hawks and doves and analysing other properties of the OSCE. The two forms of statistics complemented each other to inform the researcher and reader about the properties of the OSCE implemented so as to make a decision regarding its quality.

5.2.1 Frequency distribution of the OSCE scores

The data which comprises of the candidates' scores in each station, subject and overall were summarised by describing the distribution of the scores using the skewness, kurtosis, identifying outliers and Z-scores.

5.2.1.1 *Interpretation of skewness*

A normal data distribution assumes a symmetrical bell-shaped curve when plotted on a frequency graph. Any asymmetrical data distribution with scores spreading out towards one direction more than the other is said to be skewed. Skewness, therefore suggests the amount and direction of skew or departure from horizontal symmetry, relative to a standard bell curve.

A perfectly symmetrical normal distribution has skewness of 0. However, this is practically unlikely in a real world data. In Table 6.1 below, none of the stations' scores distribution was perfectly symmetrical. The mode, median and mean is not the same in a skewed data. In a positively skewed distribution, the right tail of the curve is longer than the left tail and the peak, shoulder and body of the curve is tilted to the left side of the graph. This suggests that the majority of the class members had low marks while a few outliers with large values, are found on the right side of the curve. These outlying scores pull the mean of the data in the positive direction, because the mean will always be located in the tail of a skewed distribution (Zygmunt & Marian 2015:2 of 8). The mean of the data will therefore be higher than the median and the mode will be the lowest value in this type of distribution. In a negatively skewed distribution, the left tail of the curve is longer than the right tail and the mean will have the smallest value, followed by the median. The mode would have the largest value (cf. Chapter 2, section 2.5.4.2). A data are highly skewed, if the skewness is <-1 or > +1; moderately skewed, if the skewness is -1 < -1/2 or +1/2 < +1; and approximately symmetrical, if the skewness is between -1/2 and +1/2 (Kim 2013:52).

In the study, all the four subjects except the overall scores had highly skewed scores (OBGYN, 3 out of 4 stations; Paediatrics, 7 out of 27; Internal Medicine, 7 out of 26 stations and Surgery, 2 out of 25 stations). All the subjects' total scores and the overall were in the approximately symmetrical category. OBGN had one out of 4 stations (25%), Paediatrics had 11 out of 27 stations (40.7%), Internal Medicine had 15 out of 26 stations (57.7%) and Surgery had 13 out of 25 stations (52%) in the approximately symmetrical category (cf. Table 5.1).

TABLE 5.1: INTERPRETATION OF SKEWNESS

GRADE OF SKEWNESS	OBGYN	PAED	INT MED	SURG	OVERALL
Highly Skewed <-1 or > +1	1-3	8,12-14,19, 23 and 26	5,6,9,14,15, 17 and 20	1 and 12	--
Frequency	75%	25.9%	26.9%	8%	
Moderately Skewed -1 < -1/2 OR +1/2 < +1	---	4,11,16-18, 20,21,24 and 25	12,19,23 and 26	2,4,6,10,11,13, 14,19,22 and 25	--
Frequency	0%	33.3%	15.4%	40%	
Approximately Symmetrical -1/2 and +1/2	4 and Total scores	1-3,5-7,9,10,15,22, 27 and Total scores	1-4,7,8,10,11, 13,16,18,21, 22, 24,25 and Total scores	3,5,7-9,15-18, 20,21,23,24 and Total scores	Overall scores
Frequency	25%	40.7%	57.7%	52%	

Discussion on skewness

The stations that were highly skewed were either too difficult or too easy for the class, depending on whether they were positively or negatively skewed. The stations in the approximately symmetrical row were close to perfection. The moderately skewed stations were of moderate difficulty or easiness.

For the stations that were significantly skewed, the mean and also the fixed university pass mark of 50% will be an inappropriate measure of central tendency. A better measure of the centre will be the median and hence the standardised pass mark for these stations (Laerd Statistics 2013:1 of 1).

The positively skewed stations suggest that the tasks were too difficult because the majority of the class had lower scores but only a few outliers in the tail of the curve had high scores. Likewise, the negatively skewed stations suggests that the tasks were too easy for the class, as the majority in the peak, shoulder and body of the curve had high scores, while the few outliers in the tail of the curve had low marks. The stations in the approximately symmetrical row were of fair difficulty.

In the study, 75%, 25.9%, 26.9% and 8% of the OBGYN, Paediatrics, Internal Medicine and Surgery stations respectively were highly skewed as shown in Table 5.1 and therefore would need to be reviewed. On the other hand, 25%, 40.7%, 57.7% and 52% of the OBGYN, Paediatrics, Internal Medicine and Surgery stations respectively and the entire total and the overall scores were close to perfection in terms of scores distribution, (cf. Table 5.1). Summing up and averaging all the stations scores per candidate per subject, somehow assisted in moving the scores close to perfect symmetry (cf. Tables 5.1).

5.2.1.2 *Interpretation of kurtosis, outliers and Z-score*

Kurtosis

Kurtosis and checking for outliers with the Box-whiskers plot were both used to identify unusually extreme values in the data. Kurtosis indicates how tall and sharp the central peak is relative to the standard bell curve. A normal distribution has kurtosis of exactly three (mesokurtic, excess kurtosis of 0). Most statistical software usually reports the excess kurtosis. A distribution with negative kurtosis or kurtosis less than 3 is platykurtic (compared to a normal distribution, its central peak is lower and broader and its tail is shorter and thinner) and it suggests low extreme scores (Laerd Statistics 2013:1 of 1). If kurtosis is positive or leptokurtic, it indicates that the data curve is tall, peaked and the tail is longer and fatter. Positive kurtosis indicates large extreme scores. The smallest possible kurtosis is 1 (or excess kurtosis of -2) and the largest is infinity. The acceptable range for kurtosis is +2/-2 (George & Mallery 2010:1 of 1). Using this guide, 50%, 7.4% and 23.1% stations in OBGYN, Paediatrics and Internal Medicine were highly kurtotic respectively as shown in Table 5.2 below.

TABLE 5.2: INTERPRETATION OF KURTOSIS

GRADE OF KURTOSIS	OBGYN	PAED	INT MED	SURG	OVERALL
Highly Kurtotic <-2 or > +2	1,3	19 and 26	5,6,14,15,17 and 21	--	
Frequency	50%	7.4%	23.1%		
Moderately kurtotic -2 ≤ -1.5 OR +1.5 ≤ +2	---	--	7 and 26	21 and 25	
Frequency	0%	0%	7.7%	8%	
Approximately Mesokurtic -1.5 >0<+1.5	2,4 and Total scores	1-18,20-25, 27 and Total scores	1-4,8-13,16,18-20, 22-25 and Total scores	1-20,22-24 and Total scores	Overall
Frequency	50%	92.6%	69.2%	92%	

Outliers

Significant outliers are scores that are outside the whiskers in the Box-Whiskers plots (Laerd Statistics 2013). Presence of outliers further strengthens the fact that the mean (and the fixed university pass mark) is an inappropriate centre for making Pass/Fail decision in this OSCE and also suggests errors in the examination process. As shown in Table 5.3 below, all the subjects including the medicine total scores had outliers. In the overall and total scores, there were no outliers.

TABLE 5.3: STATIONS WITH OUTLIERS

OUTLIERS	OBGYN	PAED	INT MED	SURG	OVERALL
STATIONS	1-3	1,2,4,8,10,18-20,23,26 and 27	3,5,6,9,14,15,17,20,24 and 26	1,2,6,10-12,15,19 and 25	--
Frequency	75%	40.7%	40%	36%	0%

Z-scores

From the raw scores, we could see extreme values. Presence of extreme scores suggests that students could obtain undeserved distinctions and fail grades for their performances. But because the scores distributions were skewed and kurtosed with outliers, it was appropriate for the scores to be standardised. Therefore, when the raw scores were converted to Z-scores, all the extreme scores shifted to the lower grades on the positive scale and higher grades on the negative scale. In the study, the best performance was 'good' (grade C) and the worst was 'poor' (grade E), (cf. Table 4.37 in Chapter 4, Section 4.9.1.1; Taylor 2014: 1 of 2; Schisterman, Whitcomb, Mumford & Platt 2009:403).

5.2.1.3 *Kurtosis, outliers and Z-score*

The skew and kurtosis are measures of shape of the scores distribution. The university fixed pass mark of 50% was based on an ideal examination process with ideal students, examiners and tests. The test scores in this ideal situation is expected to assume a normal distribution, where the centre (mean, median and mode are equal) is 50% (cf. Chapter 2, Section 2.5.4.2). In the above Sections, several evidences showed that this ideal situation was not experienced in the OSCEs conducted in this study and that the examination process may be faulted with errors probably from the examiners, students or the test itself. Actually, the ideal examination is practically almost impossible in the real world (Yusoff & Bee 2012:3 of 17).

The evidences from the above Sections that support non-normalcy in these OSCEs include skewed and kurtosed station scores distributions with outliers. Therefore, it may be inaccurate to make the pass/fail decisions based on the mean or the fixed university pass mark in a test with non-normal scores distribution. In a significantly non-normal examination situation, the median or the standardised pass mark is more accurate for use to make pass/fail decisions (cf. Chapter 2, section 2.5.4.2).

The presence of outliers in the raw scores represents extreme values which lie either in the distinction (excellent) or fail grades. However, after converting the raw scores to z-scores, these extreme scores were normalised and the best and worst performances were shifted to the lower and upper grades respectively (cf. Chapter 4.9.1.1, Table 4.37; Schisterman *et al.* (2009:403). Moreover, the Z-grades corresponded to different raw score ranges in each subject, suggesting that reporting students' performances in examinations using fixed university grades based on the raw marks from non-normal examination processes may also be inaccurate (cf. Chapter 2, section 2.5.4.1; Ganesh, Sachin & Mohini 2012:439).

5.2.2 Measures of central tendencies of the OSCE stations

The centre of the candidates' scores in each station, subject and overall were located using the sum, mean, median, mode and standard setting.

The mean is the most commonly used measure of central tendencies in a normal data distribution. The mode is the least utilised because of its weaknesses. However, where the data are significantly skewed and kurtosed with outliers, the median is the most appropriate measure to describe the centre of the non-normal data and make pass/fail decision, because the mean is the most sensitive to outliers, while the median is the most stable. The computation of the median and standard pass marks are similar, therefore the values of the median were closer to the standardised pass marks than to the mean in the study (cf. section 2.5.4.2). Moreover, it was observed in this study that the standardised pass marks were closer to the fixed university pass mark of 50%, than any of the three measures of central tendencies (mean, median and mode), (cf. Chapter 4, Section 4.9.1.2, Table 4.38). In a normal distribution, all the three measures of central tendencies are equal, at least the mean is always equal to the median (Avijiit & Nithya, 2016:18). In this study, out of the 82 stations, none of these measures were equal except in two stations: station 12 of Internal Medicine at 1.0 and station 19 of Surgery at 6.0 (cf. section 4.7.1.2, Table 4.22, section 4.8.1.2, Table 4.30).

In a positively skewed station, the mean is the highest followed by the median and the least is the mode. In a negatively skewed station, the mode is the highest followed by the median and the least is the mean. Of note is that the median remained the most stable of the three measures of central tendencies whether negatively or positively skewed (Ganesh, Sachin & Mohini 2012:442). This further supports the use of the median or better still the standard setting (both have similar computation) for determining the pass mark, rather than the mean or the fixed university pass mark.

5.2.3 Measures of variability of the OSCE stations

The measures of variability within or between the sample(s), quantifies the dispersion or spread of individual score/mean from the sample or population mean and reflect the quality of the data. They quantify the difficulty or easiness and the errors of a test. These measures of variability are used together with the measures of central tendencies, to provide an overall description of a set of data and suggests how well the mean represents the data. The variability measure of a data allows the reliability of the measures of central tendencies, to be judged. The higher the variability, the less representative the mean is in describing the data. Large spread indicates that there are probably large differences between individual scores while small spread indicates similar scores (Ganesh, Sachin & Mohini 2012:440).

Interpretation of variation depends on the purpose of the examination. In a criterion-referenced test, small variation is desirable, however in a norm-referenced test, large spread reflecting the different levels of the candidates' abilities so that ranking of the students' performances can be easily appreciated is desired. High stakes or promotion examinations are usually criterion-based. Large variations also might suggest high level of errors but difficult test. Small variations might suggest fewer errors but too easy test. When the variation is small, the scores are tightly clustered around the mean, in this case there may be a high rate of false rejection (candidates failing when actually they passed) because of the tight scoring. On the other hand, if the variation is large, the scoring may have been too lenient (high rate of false acceptances), whereby candidates passed when actually they failed.

The following measures of variability were used (cf. Chapter 2, section 2.5.4.3): range, standard deviation, standard error of the mean, 95% confidence interval of the mean, coefficient of variation, homogeneity test, ANOVA and G-studies.

The range is neither sensitive nor stable. The range is sensitive to extreme values, sample size and does not measure the variance of the data between the maximum and the minimum values. The standard deviation determines the shape of the data distribution. However, like the mean, it is sensitive to outliers and therefore may not be a very reliable measurement of variation of skewed data even though it is the most commonly used measure of variation. Standard error of the mean (SEM) judges the reliability of the sample mean by quantifying its distance from the true population mean. Standard deviation quantifies the variability within each sample, while standard error of the mean measures the variability between the samples. For the mean to be a good representative of a data, it has to be greater than 2.5 times the SD or 10 times the SEM. If the SD is higher than the mean, then the mean is certainly not a good representative of the data (cf. Chapter 2, section 2.5.4.3). Below in Table 5.4, is the analysis of the relationship between the mean and the standard deviation at each of the stations. From the table, 72% of the Surgery stations have means that were not representative of their scores. Particular attention should be paid to stations 8, 14, 17 and 19 (in Paediatrics); 12 and 15 in Internal Medicine and 22 in Surgery, whose standard deviation are actually larger than their means.

TABLE 5.4: STATIONS WHOSE MEAN IS BELOW 2.5*STANDARD DEVIATION

SUBJECTS	STATIONS	%
OBGYN	3	25
Paediatrics	6,8-17,19,20,24 and 25	55.6
Internal Medicine	7,8,11-16,18,19 and 23	42.3
Surgery	2-6,8,10-15 and 17-22	72

The coefficient of variation (CV) is an absolute measure of variation and therefore the most reliable measure of variation. Large variation is CV≥10% while small variation is CV≤5% (Avijiit & Nithya 2016:16). The report in Chapter 4 revealed that the coefficient of variation in almost all of the stations was high, (i.e. above 10%). CV was extremely high (that is above 100%) in four Paediatrics stations (8,14,17 & 19), 2 medicine stations (12 & 15) and in 1 Surgery station (22). In a high stake, criterion-referenced test, this high variation generally indicates a high level of errors, poor quality OSCE and training methods (cf. Chapter 4, section 4.9.1.3, Table 4.39).

Further proof to show that the OSCE administered differed from the ideal university examination is presented in Table 5.5 below. The stations that are missing on the table do not contain the university pass mark in their 95% confidence interval limits and therefore their scores were significantly different from the scores expected from an ideal examination (cf. Chapter 2, section 2.5.4.3).

TABLE 5.5: STATIONS WHOSE 95% CI* CONTAIN THE UNIVERSITY PASS MARK

SUBJECTS	STATIONS	%
OBGYN	2, 4 and Total	50
Paediatrics	4,5,7,9,10,11,15,20-22,24,25 and Total	44.4
Internal Medicine	1,4,8,13,16,18,19 and 23	30.8
Surgery	2-6, 8-15 and 17-25	80

CI*: Confidence Interval

The SEM of all the stations was close to 0 except in the OBGYN station 4, where it was very high at 1.1. This suggests that this station's scores distribution and its mean were probably very different from the rest of the stations in that subject (cf. Chapter 4, section 4.5.1.3, Table 4.7).

With ANOVA, the means of the stations were significantly different from each other. However, the mean of the OBGYN at station 3 was consistently and significantly higher than the mean of the rest of the stations. This strongly indicates that the examiner in that station was probably too lenient. Moreover, the variance between the stations was above 30% in each of the subjects except Surgery. The in-between variance in OBGYN was 59.6%, Paediatrics was 39.3% and Internal Medicine, 47.4% but Surgery was 24.1%. The in-between variance could be due to errors from factors external to the student such as the examiners, test and environment that could influence the students' scores. This significantly high between-station-variance suggests that these external factors apart from the differences in students' abilities are present and strong enough to influence the scores (cf. section 4.9.1.3, Table 4.40). However, this adverse effect was not seen in the overall scores, where the variance between the subjects was much less than 30% at 14.1% (cf. section 4.9.1.3, Table 4.40). The homogeneity test results indicate that the tasks in all the stations were significantly heterogeneous. However, the subject analysis suggests similarity in the construct of the tasks in each subject (Hae, 2014:2 of 6; cf. section 4.9.1.3, Table 4.42).

For the G-studies, ideally, the majority of the variances in the test should come from the students and the errors from the examiners should be very minimal (<5%). However, in this study, the errors from the examiners were significantly high (especially in OBGYN-89.2%, cf. Chapter 4, section 4.10.3, Table 4.46) and higher than those from the students. The peculiar situation in OBGYN could be explained by the fact that all the 4 internal examiners were part-timers. There was only one specialist amongst the internal examiners, the rest of them were medical offices with only MBCHB degree. These internal examiners were also responsible for the teaching and training of the students in OBGYN. The low academic level of the internal examiners may be accompanied by poor skill in teaching, clinical training, assessment and scoring of the students and could therefore impact negatively on the OSCE scores. In the overall, the examiners' variances were high, above 5%, but less than the students' variance which was at 54.3% (cf. section 4.9.1.3, Table 4.41). There were a lot of interactions between the students and the examiners. The student-examiner interaction was highest in Surgery (28%) which had the highest number of junior examiners, followed by Paediatrics (26.5%) (cf. Chapter 4, section 4.10.3, Table 4.46). Considering that Paediatrics had the least number of examiners, the level of errors from student-examiners' interaction in Paediatrics was very significant. The G-coefficients were moderately low (cf. Chapter 4, section 4.10.3, Table 4.46). This is evidence to suggest that the facets external to the students (e.g. examiners, environment etc.) were strong enough to influence the students' marks (Pugh, Hamstra, Wood & Bordage 2014:11 of 16; cf. Chapter 2, section 2.5.4.3).

5.2.4 OSCE station analysis

The station analysis tests measures the quality of the tasks in the stations. The properties of each OSCE station was analysed with the following tests: IDI (Difficulty Index), *d* (Discrimination Index), statistical significance between means of pass and fail students and failure rate.

IDI measures the easiness of the tasks in a station. Good IDI is between 0.3-0.8 for a norm-referenced test, to show the different capabilities of each student. However for a criterion referenced test, IDI up to 0.95 is desirable, to show that most of the students have gained mastery over the tasks in the station. This final OSCE is a criterion referenced test. IDI is the reverse of the failure rate. Good discrimination is the ability of the tasks in the stations to differentiate between the academically strong and weak students. Too easy or too difficult tests have poor discriminative power. According to Tavakol and Dennick (2011b:453), good *d* is between 0.3-0.5. If *d* is 0 (in other words, all have passed), it suggests poor or good discrimination. A good station might have '0'*d*, if the task/s in the station is/are basic essential components of the subject that all students (strong or weak) should have mastered, for example cardiothoracic resuscitation. Hamza, Gominda and Zubair (2014:227) graded *d* into poor (≤0.2), moderate (>0.2≤0.35) and high (>0.35). The statistical significance further confirms the discriminative power of a station by measuring the difference in the means of the students who passed versus those who failed the tests in each station (cf. Chapter 2, section 2.5.4.4). IDI (failure rate), *d* and the statistical significance are used together to describe the quality of the tasks in the stations.

From Table 5.6 below, the strong stations were those with good or high IDI, statistical significance and *d.* The poor stations were those with bad or poor IDI, *d* and statistical significance. Stations with moderate quality tasks are those with good IDI and statistical significance and moderate *d.* The poor stations included those with negative *d* (cf. Chapter 2, section 2.5.4.4; Table 5.6 below)*,* suggesting that the good students were failing the station, while the weak ones were passing it, hence the tasks in these stations were very weak at discriminating (cf. section 2.5.4.4). The stations with negative *d* were station 1 in OBGYN, stations 6 and 7 in Paediatrics and stations 7 and 22 in Surgery. The strong stations should be kept in the OSCE bank for future use, the moderately strong stations should be reviewed, while the poor stations should be discarded and replaced with tasks of good quality. In this study, none of the OBGYN stations had sufficiently good properties, (cf. Chapter 4, section 4.5.1.4.; Table 4.10; Table 5.6 below). There were 8 (29.6%) strong stations in Paediatrics (1,4,5,12,13,20,21 & 26). In Internal Medicine, only 2 (7.7%) stations (8 and 18) fall in the category of strong stations. Surgery had 14 (56%) strong stations (2,5,8-13,15-20). Moreover, no student attempted the tasks in station 23, possibly because the students were not familiar with the topic. Perhaps they were not taught, and if they were taught, it is possible that none understood the topic to the extent that all avoided to attempt it. If the latter case holds, then the teaching methods need to be investigated.

TABLE 5.6: INTERPRETATION OF STATIONS ANALYSIS

IDI*	POOR *d*	MODERATE *d*	HIGH *d*	POOR SS**	GOOD SS**	TOTAL
	OBGYN					
Good (0.3-0.9)	4	2 and Total	--	2	4 and Total	2 stations
Bad	1,3			1,3		2 stations
	PAEDIATRICS					
Good (0.3-0.9)	6,7,10,17, 22	Total	1,4,5,8,9,11,12, 13, 15,20,21,24-26	6,7,8,9,10, 11,15,17,22, 24,25 Total	1,4,5,12, 1320, 21,26	19 stations
Bad	2,3,14,16, 18,19,23,27			3,14,16,19, 27	2,18,23	8 stations
	INTERNAL MEDICINE					
Good (0.3-0.9)	1,3,4,9,10, 13,17	11,16,19, 22	8,18,20,23 Total	1,3,17,20,23	4,8-11,13, 16,18,19, 22, Total	15 stations
Bad	2,5-7,12,15,21 24-26	14		5,6,14,25	2,7,12,15, 21, 24,26	11 stations
	SURGERY					
Good (0.3-0.9)	3,4,6,7,14, 16,21-25***Total		2,5,8-13,15-20		2-25***Total	23 stations
Bad	1				1	1 station

*IDI: Item Difficulty Index, **SS: Statistical significance, ***Station 23 is excluded from the table.

5.2.5 Reliability checks

The reliability tests indirectly measure the amount of errors in a station and the internal consistency of the tasks in each station. The reliability tests used in this study were: α-correlation, α-deleted, Pearson correlation (*r*) and *r^2.*

The Pearson correlation measures the relationship or the agreement of the station or subject scores with the total or overall scores respectively. The Pearson correlation also is superior as a discrimination index to *d,* as it takes into account the scores of all the candidates, while *d* only considers the scores of the candidates in the upper and lower third (Hamza *et al.* 2014:228). Total alpha measures the internal consistency while r^2 measures the importance of each station. The closer the correlations are to 1, the better the reliability. The stations were categorised into strong, moderately strong, low and weak based on the correlations guideline (Hemphill 2003:78; cf. section 4.9.1.5). The α-deleted reliability value of a station should decrease below the correlation value, if the reliability is good. In stations with poor reliability, the α-deleted improves and if it improves above the total α, the questions in the station are redundant or are being repeatedly used, though in a different format (cf. section 2.5.4.5).

5.2.5.1 *Interpretation and discussion of the Pearson's correlations*

In Table 5.7 below, the stations with strong correlations and useful r^2 were the most contributory in the OSCEs. Only 3 (3.7%) stations were strong and useful. The stations with weak *r* and non-useful r^2 especially those with negative correlations need urgent review, because they contributed very little to the total scores (cf. section 2.5.4.5).

TABLE 5.7: INTERPRETATION OF CORRELATION(R) AND USEFULNESS (R^2)
(Table continue on next page)

CORRELATIONS	USEFUL STATIONS	NOT USEFUL STATIONS	TOTAL
	OBGYN		
Strong	4	----	1 station
Low	----	3	1 station
Weak	-----	1 and 2	2 stations
Total	**1 station**	**3 stations**	**4 stations**
	Paediatrics		
Moderate	---	1,5,8,9,12,13,20 & 21	8 stations
Low	---	2,4,11,17-19,24 & 26	8 stations
Weak	---	3*,6,7,10,14-16,22,23,25 and 27*	11 stations
Total	**---**	**27 stations**	**27 stations**
	Internal Medicine		
Moderate	**---**	8	1 station
Low	**---**	1,2,4,7,10,11,13,14,1819,21-24 and 26	15 stations
Weak	**----**	3*,5*,6,9,12,15-17*,20 & 25	10stations
	----	**26 stations**	
Total			**26 stations**
	Surgery		
Strong	12 and 14	----	2 stations

Moderate	---	13 and15-20	7 stations
Low	---	1,2,5,6,8-11,21 & 24	10 stations
Weak	----	3,4,7*,22* & 25	5 stations
Total	**2 stations**	**22 stations**	**24 stations**
Grand Total	**3stations**	**78 stations**	**81stations**

(*negative correlation). No student attempted station 23 in Surgery of the Main study.

5.2.5.2 *Interpretation and discussion of the Alpha correlations*

The stations with good internal consistency and therefore strong reliability are the ones with strong *α* correlations and decreased *α*–deleted values (cf. section 2.5.4.5). In Table 5.8 below, only 3 (3.7%) of the stations had strong *α* correlations and decreased *α*–deleted values. The stations under the redundant columns (especially the weak stations with negative correlations) have tasks that have been repeatedly used and therefore need to be changed. OBGYN stations had the worst reliability coefficient (cf. Table 5.8; Al-osail, Al-osail, Al-ghamdi, Al-hawas, Al-bahussain & Aldajani 2014:195).

TABLE 5.8: INTERPRETATION OF CORRELATION (α) AND α- DELETED
(Table continue on next page)

***α*–CORRELATIONS**	**DECREASED**	**IMPROVED**	**Redundant**	**Total**
		OBGYN *(α=0.008)*		
Weak	3&4	----	1*&2*	4stns
Total	2 stns	---	2 stns	4 stns
		Paediatrics *(α=0.6)*		
Moderate	9&12	1,13,18&21	---	6 stns
Low	---	4,5,8,20,24&26	---	6 stns
Weak	---	19&22	2,3*,6*,7*,10,11, 14*,15,16*,17, 23,25,&27*	15 stns
Total	2 stns	12 stns	13 stns	27 stns
		Medicine ***(**α=0.36)*		
Moderate	8&22	---	----	2 stns
Low	10	2&7	---	3 stns
Weak	---	1,4,11-15,18,21,24-26	3*,5*,6*,9*,16*,17*,19,20 & 23*	21 stns
Total	3 stns	14 stns	9 stns	26 stns
		Surgery *(α=0.76)*		
Strong	12&14	----	---	2 stns
Moderate	---	13	---	1 stn
Low	---	5,8-11,15-20&24	1	13 stns
Weak	----	----	2-4,6,7*,21,22*& 25	8 stns
Total	**2 stn**	**13 stns**	**9 stns**	**24 stns**
Grand Total	**9 stns**	**39 stns**	**33 stns**	**81 stns**

(*negative correlation). No student attempted station 23 in Surgery.

The total alpha and G-coefficients were low-moderate in this final OSCE, Table 5.9 below. This is contrary to the general expectation that final examinations are usually the most standard examination in the university, as much more efforts are usually invested in an exit examination than in any other examination in the university, and therefore the reliability of an exit examination should be within the acceptable range.

TABLE 5.9: COMPILATION OF THE RELIABILITY ESTIMATES

SUBJECTS	α-COEFFICIENT	G-COEFFICIENT
OBGYN	*0.0*	0.0
PAED	*0.6*	0.2
INT MED	*0.4*	0.1
SURGERY	*0.8*	0.4
OVERALL	0.2	0.8

5.2.6 Hawks and Doves

An examiner who scores to the extreme could either be identified as a Hawk or a Dove. The Hawk has a significantly lower mean than the rest of the stations in the same subject, small variability (CV below 5%, with a lot of false rejections), poor correlations and importance. The Dove has a significantly higher mean than the rest of the stations in the same subject, small variability (CV below 5%, with a lot of false acceptations), poor correlations and importance. The Dove has been too lenient while the Hawk has been too strict with the scoring of the students' performances (cf. section 2.5.4.7). Amongst the Hawks and Doves displayed in Table 5.10 below, OBGYN station 3 was particularly too lenient and therefore may need to be reviewed.

TABLE 5.10: HAWKS AND DOVES

SUBJECTS	HAWKS	DOVES
OBGYN	----	3
PAED	----	18 &19
MED	-----	5
SURGERY	*22*	1

5.2.7 Other attributes of the OSCEs

Overall, the variability was high, students' errors (G-coefficient=0.8) and the difficulty index (IDI =0.67) were moderate in this study and suggests that the OSCE was generally tough from the students' angle. Overall, errors from the examiners were low, less than 30%. However, the internal consistency of the OSCE was weak at 0.24 (cf. section 4.9.1.5). The gender means, quality of the borderline test in determining pass marks and the effect of junior examiners on the G-coefficients were also assessed.

5.2.7.1 *Correlations between manned and written OSCE stations' scores*

The correlations between the manned and written stations' scores were high in all the subjects except in OBGYN, where there was no correlation at all. The written stations mean was lower in the written stations, and suggests poor teaching methods in OBGYN (cf. section 4.10.1, Table 4.44).

5.2.7.2 *Gender and means*

The female candidates' mean was significantly higher than the male candidates' mean in Paediatrics. Of note, the level of interaction between the examiners and students was second highest in Paediatrics which had one of the least number of examiners. However, there were equal number of male and female examiners, the sample size is small and therefore little importance can be attached to this finding (cf. section 4.10.1, Table 4.46).

5.2.7.3 *Effect of junior examiners on the G-coefficients*

There appears to be no significant relationship between the levels of examiners errors, student-examiners' interactions, G-coefficients on one side and the number of Junior examiners that participated in the OSCEs on the other hand (cf. section 4.10.1, Table 4.46). The positive and high correlation between the number of junior examiners and the G-Coefficient, suggests that the higher the number of junior examiners, the higher the G-Coefficient and therefore the errors in the tests comes only from the students. Also the number of junior examiners was inversely related to the level of examiners' errors. However, these relationships were not significant and the contribution of the junior examiners to the differences seen in the levels of examiners errors, the student-examiners interactions and the G-Coefficient (student variability) were low to moderate (cf. section 4.10.1, Table 4.46).

5.2.7.4 *Quality of the borderline test*

The borderline test was used to determine the standardised pass marks. The sensitivity, specificity and predictive values of this test were determined in comparison to the university pass mark. The sensitivity of the borderline test (i.e. its ability to identify students that have passed correctly) and its predictive value for failing (i.e. the level of accuracy of this test to identify failing students) was moderate in this study. However, its specificity (i.e. its ability to identify failing students) and predictive value for passing were very high. This shows that the test is more reliable as a tool to guide the examiners in determining those who have truly passed, while it is less reliable in determining students that are failing. Hence, the university pass mark can be used as a screening test while the borderline test can be used to confirm those who have truly passed (cf. Chapter 4, Section 4.10.4; Table 4.47).

5.3 SUMMARY

The following is the summary of the properties of the OSCE that was conducted in July 2015 as reported in this study.

TABLE 5.11: SUMMARY OF THE PROPERTIES OF THE OSCEs
(Table continue on next page)

FINDINGS	OUT OF 82 STATIONS		
Non-normal distribution of scores: unusual extreme scores			
Highly skewed station (cf. Table 5.1)	19(23.5%)		
Highly kurtosed stations (cf. Table 5.2)	10(12.3%)		
Outliers (cf. Table 5.3)	33(40.7%)		
Performances (cf. Section 4.9.1.1).	Poor for a criterion referenced test		
z-grade ranges (cf. Section 4.9.1.1).	C-F		
G-studies (cf. Section 4.10.3, Table 4.46)	Weak-Low G-coefficient, Very high examiners' errors and student-examiner interactions		
Subjects where the examiners errors superseded the students' errors	4(100%)		
Item analysis	Not ideal for a criterion ref test		
Stations and item analysis (cf. Table 5.6)	Good (Keep)	Not sufficient (Review)	Bad (Discard)
OBGYN	----	2&4 (50%)	1&3 (50%)
Paediatrics	8 (29.6%):1,4,5,12, 13,20,21 & 26	14(51.9%) 2,6-11,15,17, 18,22-25	5(18.5%) 3,14,16, 19 & 27
Medicine	2(7.7%) 8 &18	21(80.8%) 1-4,7,9-17,19-24	3(11.5%) 5,6&25
Surgery	14 (56%) 2,5,8-13,15-20	10(40%) 3,4,6,7,14, 21-25,23	1(4%) 1
Correlations(*r*), Table 5.7	Very few stations with strong correlations and usefulness		
Non-useful stations with low-weak correlations (*r*), Table 5.7	62(76.5%)		

Reliability estimates (alpha), Table 5.9	Weak(OBGYN)-High(Surgery)
Redundant stations with low-weak alpha (cf. Table 5.9)	33(40.7%)
Hawks and Doves, Table 5.10	6(7.4%)
Significant gender means differences (cf. Table 4.45).	Paediatrics (Female mean>Male mean)
OVERALL	
Variability (cf. Section 4.9.1.3)	Very high
Errors from students (cf. Section 4.9.1.3, Table 4.41)	Moderately high
Examiners' errors (cf. Section 4.9.1.3, Table 4.41)	Low
IDI (cf. Table 5.6)	Moderate
Internal Consistency (cf. Table 4.43)	Weak

Of note is that, the presence of junior examiners in the OSCEs did not significantly affect the examiners' errors, the student-examiners interaction nor the G-coefficient (cf. Chapter 4, section 4.10.3, Table 4.46). Borderline test for standardising pass mark had high specificity and predictive value for passing but moderate-low sensitivity and predictive value for failing. Station 23 tasks of the main study Surgery OSCE was not attempted by any student. Students were unfamiliar with the topic.

5.4 CONCLUSION

In this chapter, it was found that the final OSCEs have not been the ideal examination set up that the university desires for a high stakes criterion referenced assessment, contrary to the general expectations, because more university efforts and resources are usually invested in the final examinations as compared to any other examinations. The summary of the psychometrics properties of these OSCEs was presented in Table 5.11.

The subsequent Chapter 6 will give the Conclusion, Recommendations and the Limitations of the study.

CHAPTER 6

CONCLUSION, RECOMMENDATIONS AND LIMITATIONS OF THE STUDY

6.1 INTRODUCTION

This chapter provides a summary of the research findings and discusses the implications of these findings for OSCEs in medical schools. In this study, research was carried out to illustrate how psychometric analysis can objectively measure and ensure quality of the OSCEs and make recommendations and guideline for the regular use of psychometric methods. The findings have revealed the understanding of these examinations in the medical schools in Tanzania, how they are conducted, and how they are scored. Psychometric analysis is a stable, objective, post-examination component of the quality assurance system that evaluates students' scores (cf. section 2.4). Making inaccurate judgement on the competency of medical trainees could place patients at risk. OSCE being resource-intensive, its delivery may be compromised in resource limited institutions. Therefore, it is important to measure both the quantity and the type of quality of the OSCEs practiced especially in resource limited institutions. The final OSCEs done in this study were summative and based on criterion or absolute referencing, where student mastery of the subject is desired rather than norm-referenced which is not recommended for undergraduate training nor for clinical competency tests (cf. section 1.1).

The aim of this chapter is to present comments and some concluding thoughts on the findings and to provide a short overview of the study. A literature review and the illustration of the application of the psychometric methods on live OSCE raw scores, were done in this study (cf. Chapters 2 & 3). This chapter commences in section 6.2 with a summary of the study findings as they are related to study objectives or questions. This is followed by a short discussion on the limitations of the study (cf. section 6.3). In turn, Section 6.4 provides recommendations and the contribution to knowledge with the provision of a user-friendly guideline for the purpose of carrying out a comprehensive psychometric analysis for the establishment of regular psychometric analysis into the quality assurance system of the school to measure and ensure the quality of the OSCEs and other types of assessments in resource limited medical schools. Finally, the pointers to further research were presented in section 6.5, an overview of the whole study is provided in Section 6.6 and then the final remark is made in section 6.7.

6.2 SUMMARY OF FINDINGS

The research was carried out and completed based on five research questions. The findings of the research served as a foundation for making recommendations and guideline for the use of psychometric analysis to measure the quality of assessments at KIU-Dar.

6.2.1 Research question 1

The research question was stated as:

What should the psychometric methods that objectively and comprehensively evaluate the OSCEs with a view to ensure quality in an institution with limited resources look like?

For research Question 1, the following objective was pursued:

Describe the available psychometric methods for the OSCE through literature review.

This was extensively addressed in Chapter 2 as Literature Review.

6.2.2 Research question 2

The research question was stated as:

How should available psychometric methods of analysis of an OSCE be objectively and comprehensibly applied with a view to ensure quality in an institution with limited resources at KIU-Dar in Dar es Salaam?

For research Question 2, the following objective was pursued:

Develop a guideline for psychometric analysis and interpretation of the OSCE scores in a medical school.

The guideline followed is summarized in section 6.5 below.

6.2.3 Research question 3

The research question was stated as:

How can psychometric methods be applied to analyse and interpret the raw scores of an OSCE at a medical school?

For research Question 3, the following objective was pursued:

Illustrate the application of the available psychometric methods on the raw scores from a live summative OSCE in a resource limited medical school at KIU-Dar in Dar es Salaam to obtain its psychometric properties and provide an appropriate interpretation of these properties.

This was extensively addressed in Chapters 3, 4 and the summary is also provided in the guideline in section 6.5 below.

6.2.4 Research question 4

The research question was stated as:

What is the quality of the OSCE currently practiced in a typical resource limited medical school at KIU-Dar in Dar es Salaam?

The following objective was pursued:

Describe the quality of the OSCE that is currently practiced in a typical resource limited medical school at KIU-Dar medical school in Dar es Salaam.

This objective was achieved in Chapters 4 and 5. The quality of the OSCEs currently practiced in KIU-Dar es Salaam was measured with 27 psychometric tests grouped into six categories in Chapter 4 and interpreted in Chapter 5. These psychometric tests described the reliability of the OSCE scores which strongly reflects the validity of the OSCE (cf. Chapter 1, section 1.1; Downing 2004:1007). The summary of the reliability of these OSCEs is discussed below.

6.2.4.1 *Reliability*

Reliability of a test is described as the precision or quality of the test in discriminating students' performances upon repetitions and when examiners are in close agreement in their ratings (Boursicot, Etheridge, Setna, Sturrock, Ker, Smee & Sambandam 2011:375). Reliability is the most significant property of an assessment. The General Medical Council of London emphasise the importance of using the reliability metrics to quality assure and improve assessment process. An unreliable test cannot be valid. Furthermore, reliability can also be subcategorized into stability (stable students' performance upon repetitions) and internal consistency (consistent score correlations with the sum of all other scores). Internal consistency for each individual station is more important than the overall internal consistency for the purpose of station analysis and revision (Norcini, Anderson, Bollela, Burch, Joa, Costa, Duvivier, Galbraith, Hays, Kent, Perrott & Roberts 2011:207).

Cronbach alpha correlations were used to measure the internal consistency and stability of the OSCE scores in this study. The number of assessors, cases and adequate test time are the most important factors determining the reliability of a test (Tavakol & Dennick 2012:e161). From several studies, a 2-6hrs OSCE with 8-12 stations will yield a reliability coefficient of between 0.41-0.88. Even the most stringent ECFMG OSCEs reports reliability coefficient of 0.64. However, Al-Naami *et al*. (2008:300) documented the stability of the OSCE in a surgical examination in King Saud University, measured using Cronbach alpha to be between 0.87-0.88. In this study, the internal consistency in the stations ranged from -0.4 to 0.8.

The overall stability of the OSCEs in this resource constrained institution was 0.24, which is weak. However, a higher stability was recorded in each of the subject OSCE, except in OBGYN. For example, stability for Surgery was 0.76, Medicine, 0.36, Paediatrics, 0.6 while OBGYN was 0.008, (cf. Chapter 5, cf. Table 5.8). These values are low compared to documented studies in the developed countries based on the reliability estimates and level of errors experienced in these assessments.

Moreover, the G-coefficient in this study was unstable with the overall G-coefficient at 0.6 but lower for individual subject. The overall G-coefficient (0.75) is comparable to other studies, however, since the students have to pass each course to graduate, the collective assessment G-coefficient cannot serve as an indicator of the quality of the OSCE. For example, Blood, Park, Lukas and Brorson (2015:1623) recorded G-coefficient of 0.64 in Chicago whereas Bajwa, Yudkowsky, Belli, Vu and Park (2016:1 of 1) documented G-coefficient of 0.45 in Geneva. Regarding station analysis (difficulty and discriminative indices), Wallenstein and Ander in Georgia (2015:121) documented a range of 53-73 difficulty index and 0.5-0.6 discrimination index. In KIU-Dar, the ranges were 0-1 for difficulty index and -0.4 to 1 for discrimination index.

The Generalizability coefficient, station analysis, internal consistency and stability were poor in this final OSCE. The weak individual G-coefficient (OBGYN:0.0; Paediatrics:0.2; Medicine:0.1 and Surgery:0.4) experienced in this study suggests that the factors external to students were strong enough to influence their score. It is expected that the final examinations would be the most organised and standard assessment in the university, considering the significance, efforts and resources invested into it. However, it was disappointing to discover, in this study, that the psychometric properties of the final OSCE was very weak (cf. Chapter 5, cf. Table 5.10).

Assessors

The goal is to have at least one assessor per station. Individual assessor bias can be reduced by the use of multiple assessors. The high level of examiners errors (51-89%) and errors from their interactions with the students (11-28%) experienced in this research were unacceptable for a criterion referenced assessment. The high examiner-student interactions experienced in this study could arise from inadequate training as insufficient time was allocated for briefing which is not the same as training. The number of examiners in this study was inadequate. Examiner: station ratio was 1:4. This could have contributed significantly to the examiners' bias/errors. Shortage of teachers / examiners is a common problem in resource limited schools.

The researcher documented medical officers (junior examiners) as part of the examiners' panel especially in surgery, which had the highest number of junior examiners. From the analysis, the junior examiners did not significantly affect to quality of the OSCEs. This was probably because these medical officers have had long years of experience in medical practice. From these findings, it would appear as if the academic level of the examiners does not influence the OSCE performance. This assumption, however, need to be taken with caution because of the small sample size used in this study.

Test time

The test time in this study, ranged from 30 minutes to 2½ hours which is not sufficient for adequate reliability (cf. section 2.5.2; cf. Steven 2011:85). The students passed through so many stations in a short time suggesting superficiality of the tasks in the stations. No rest station was provided amongst these multiple stations, bringing in the issue of student fatigue as a source of error, which may have contributed to the 54.3% student variance, where so many stations were used.

Cases

The cases were few, compared to the number of stations created. Most of the stations were converted to written, which is not appropriate for a clinical examination (cf. section 4.1, cf. Table 4.1).

Scoring

Scoring also affects the reliability of a test. The significant gender differences noted in OBGYN and paediatrics together with their significantly high examiners' errors and interactions, suggests that further investigations into the scoring system of these two subjects need to be carried out (cf. Chapter 4, cf. Table 4.45). The written stations were co-marked by the medical officers and therefore the term 'hawk' and 'dove' as it applies to only one examiner may not have been appropriately used in this study. Hence, in the stations where 'hawks' were identified, it is better to state that they were too strictly scored and where 'doves' were found, the stations were too leniently marked. In this research, there was no global scoring because the examiners claim they were not familiar with it and the MoH checklist that was used, did not incorporate global scoring. The researcher could not therefore assess the quality of the checklist versus the global scores. Global scoring has been found to be more reliable than the checklists (Kim, Neilipovitz, Cardinal & Chiu 2009:6). This problem of omission could have been taken care of, if the checklists were designed to also contain the global rating in its simplified form. It is better to determine the pass mark from both the checklist and global scores using the borderline regression method (Boursicot *et al*. 2007:1024).

Standard setting

In absolute referencing, the cut-off score is defined, thereby identifying candidates who are competent and eligible for progression. The method used to determine the cut-off score is critical in this set up. In this university, where this study was carried out, a fixed university cut-off score of 50% has been used since inception to make decision of pass/fail. However, this fixed 50% is based on an ideal examination set-up and process. From the researcher's observation in this study, the OSCEs administered were not ideal, (cf. section 5.3, cf. Table 5.12). Therefore, it will not be appropriate to use the mean nor the fixed university score, to determine the cut-off score.

Tavakol and Doody (2015:38) stressed that medical educators must ensure the reasonableness of the pass mark and the quality of all test items to avoid unfair failing or passing of students that might result in students' complaints, legal actions and political issues (Tavakol & Doody 2015:38). Boursicot *et al.* (2007:1024) recommended the use of the borderline group or borderline regression method or contrasting groups for setting standards for decision making in criterion referenced tests. These 'gold standard' methods were developed specifically for the OSCEs by the Medical Council of Canada (Boursicot *et al.* 2007:1024). The border-line method (if there is no global score) used in this study is a better method for determining the cut-off score (Boursicot *et al.* 2007:1024). The borderline method had high specificity and predictive value for passing candidates in this study.

Variance

The generally high variance observed in this study is not desirable for a criterion-referenced OSCE (cf. Chapter 5, cf. Table 5.12). A narrow variance is desirable in absolute-referenced assessments.

The only variance acceptable in the OSCE, especially for high stake examinations, permitted to influence the students' performances or scores is that coming from their different levels of ability. An estimate of examiners errors of 12% was documented in Britain and the USA and 10-17% in the Medical Council of Canada (MCC) examinations (Bartman *et al*. 2013:28). Boulet *et al*. (2003:28) reported 10.9% in the ECFMG Clinical Skills Assessment (CSA) and Roberts *et al*. (2010:690), 8.9% in Australia. However, Harasym *et al*. (2008:617) claimed examiner variance that was more than four times the examinee variance in the University of Calgary, Canada. This study recorded examiners' variance in the range of 50.7-89.2%, which in several instances overwhelmed that of the students in the four subjects (especially in OBGYN) under investigation. The test itself and student-examiner interactions also contributed to the variances in this study. According to Bartman *et al*. (2013:28), examiners errors could be due to leniency, inconsistency, halo effect and restriction of ranges.

Because the variances in this study were high together with the very high level of examiners errors, poor internal consistencies, stability and failure of the measures of central tendencies in some stations to follow expected pattern, excessive leniency and inconsistencies could explain the errors from the examiners

Student's performances (Z-scores)

This research showed wide gaps or differences or variation in the levels of mastery of each subject by the students from very poor to very good. No candidate had distinction after normalising the scores to remove outliers. This also shows some weakness in the teaching/training methods used in this resource limited school. For an absolute referenced assessment, it is desired that the candidates be taught and learn properly with the goal of achieving uniform mastery of the subjects across all the students.

Currently, this school uses the raw scores to grade students, but from what was observed in this research, this is erroneous as this might lead to the award of undeserved distinction or failure. Also, it should be noted that the score ranges matching each of the grades were different in each station and subject. Therefore, using a fixed score ranges for each grade by the university is wrong (cf. section 5.2.1.2).

Other aspects of the reliability of the OSCEs

The researcher by measuring the reliability was indirectly investigating the validity of these OSCEs (cf. Chapter 1, section 1.1). Hence, it is worth mentioning some aspect of validity observed in the OSCE done in KIU-Dar. Validity is the extent to which a test measures what it is supposed to measure (Downing 2004:1007).

The content (comprehensiveness) validity of a test is provided by a blueprint which is a framework for content areas of the test (Boursicot & Trudie 2005:16; Boursicot *et al*. 2011: 375). The OSCE was to test various areas of competencies such as:

- Knowledge;
- History-taking skills;
- Doctor-patient communication;
- Professionalism skills;
- Physical examination skills;
- Clinical reasoning skills;
- Practical and technical skills;
- Ethical behaviours;
- Relationship with colleagues;
- Management skills; and
- Problem solving skills.

(cf. Chapter 1.1)

There was no blueprint for the OSCEs carried out in this research. Moreover, less than 20% of the stations were manned with real patients. Majority of the stations were converted to written format, then co-marked by very few examiners and this could have contributed to the huge amount of examiners' error experienced in this study. Also, from the researcher's observation, the manned stations available were designed to test for history taking, some few systemic physical examination and procedural. None of the manned stations tested on the central nervous system, neonatology, interpretation of laboratory results and communication skills. OBGYN had only two manned stations, hence the coverage of the learning outcomes in the OSCE implemented was most likely small (cf. section 4.1, cf. Table 4.2).

There was also no variability in the length of the stations to suit the objectives and the different nature of the tasks in each of the stations and this might weaken the validity of the OSCE. All the stations were five minutes long, including those for history taking, which should have taken longer to achieve a meaningful evaluation. All the OSCEs per subject in each study were done in a session per day. Paediatrics, Internal Medicine and Surgery used too many stations in a session without any rest station in between. OBGYN on the other hand used too few stations. Station 23 of Surgery OSCE should be excluded as none of the students attempted the station because they were not familiar with the topic of the station. This point to a weakness in the teaching method employed in the school and inadequate clinical exposure. This has a negative impact on the validity. Moreover, the correlation between the written and manned station scores was high in all the subjects except OBGYN, where, there was no correlation at all (cf. section 4.10.1, cf. Table 4.44).

The accuracy or concurrent validity of the clinical stations with the MCQ and the written stations were not measured in this study partly due to difficulty in accessing results. It is better to compare the OSCE scores with scores from other methods of clinical assessments. This could not be done here because OSCE is currently the only clinical assessment the candidates are subjected to in this school.

The construct validity is the ability of the test to differentiate clearly between the strong and the weak students or students with different levels of competencies. The station analysis in this study showed that the best station analysis report was in the internal medicine OSCE. The other subjects had poor discriminating power and difficulty index for a criterion referenced assessment. The alpha and Pearson's correlations also provides a good reflection of the construct validity and the internal consistency of a test (Hamza, Gominda & Zubair 2014:227). From Tables 5.8 and 5.9 (cf. Chapter 5.2.4.1.) only 3 and 9 out of the 82 stations (3.7%, 11%) paraded in the final OSCE met the criteria of strong and useful according to Pearson's and alpha correlation respectively. The global score is an additional evidence of construct validity, however, this was not provided in this OSCE.

Feasibility of the OSCE has improved in other studies, however in this study, feasibility is still a challenge because of very limited qualified examiners, constraint in finances and facility resources. The most obvious lack is the absence of a university teaching hospital for this institution. A regional referral hospital, which was about 30 minutes' drive away from the university, was used for both the EYE and the final OSCEs. The university had to rent the venue and obtain consent to conduct the OSCEs in advance from the hospital administrators under their regulations. This venue was constructed and functionally targeted more towards patient care rather than for training and assessments. It was not very conducive for the OSCEs. Feasibility will be better in the future, when the university teaching hospital is completed. Currently, there is no medical education department in the school to directly supervise medical training. If a medical education department is in place, there would be no need to obtain consent from the departments to evaluate their examinations, and the challenges of obtaining the results for analysis, will be much reduced.

In this study, funds were provided by the university to convey the students to the venue of the OSCEs, payments were made to the hospitals for the venue, consumables used, hospital administrators, examiners and patients. The cost of this OSCE was about 25,000/= (US $12) per student. This current cost is very low compared to the minimum of $200 per students in the US in 2012 and the cost of comprehensive OSCE at $1,080 per student in Quebec medical school for example (Turner & Dankoski 2008:574). University staff motivation was poor in the OSCE implemented in this study. No standardized patient was used. A structured simulated patient training program is lacking in the institution. There were few examiners and real cases to satisfy the large number of candidates and stations. Interviewing and examining actual patients by the increasingly large number of medical students is becoming an ethical dilemma in this medical school. Incorporating all these missing items will swell the cost of implementing OSCEs in this institution in the future.

The OSCEs implemented showed that moderate learning has taken place, though not satisfactory for a referenced based assessment, where high grade performance is desired for every candidate to show their mastery. The range of performance is very wide (from very good to fail) showing different levels of understanding of the subject amongst the students. This heterogeneous learning status amongst the students, suggests inadequate teaching or clinical exposures or slow learners. Likewise, station 23 of surgery department in the study, where none of the students attempted the tasks, could have had a negative impact on the psychometric properties of these OSCEs, training and student performance. The findings and recommendations in this study, will guide the departments, trainers and students on how to improve subsequent OSCEs for better training and learning.

OSCE has been well accepted in this medical school by all stakeholders (students, lecturers and policy makers) as it is less time consuming during administration especially when handling large number of students as compared to the traditional method of clinical assessment. There are also other benefits of the OSCE which has been documented in other studies. However, we need to objectively measure the level of acceptability in this academic community in future research.

In the US, UK and other countries, there are no standard operating procedures for running psychometric analysis of the OSCEs in the medical schools as at the time of this study. Hence, there will always be institutional variation on how the OSCEs are delivered and evaluated. No valid single method of assessment exists. The best approach is to use multiple assessment tools longitudinally to form a more holistic opinion of individual candidate's level of clinical competency and overcome the inadequacies of individual assessment method (Brannick *et al*. 2011:1181).

6.2.4.2 *Other deficiencies in the OSCE implemented*

Below are other challenges experienced in this OSCE.

- The university is yet to have its own teaching hospital. The venue of the OSCEs was in a referral hospital purposed for patient care rather for teaching. Therefore, the space and venue were neither adequate nor appropriate.
- Depending on only OSCE to test clinical skills.
- The examiners were few.
- The stations were too few in the pilot study and too many in the main study.
- There was no blueprint.
- There was no simulated/standardised patients.
- There was no global scoring.
- Lack of a proper scores grading system.
- Lack of standardised method of determining the cut-off scores.
- The length of the station was uniform throughout, therefore the length of the stations

may not be fitted to the tasks in the station.

- Insufficient training of examiners on OSCEs and its scoring.
- Staff motivation was poor.

However, in several places in the interpretation of the report in Chapter 5, at the time of merging the stations' scores, several of these adverse effects tend to cancel each other out (cf. Chapter 2.6). Where the total scores remain in the red, then the offending stations in that subject seriously need to be reviewed. But this 'better looking data' as a result of merging cannot totally obfuscate the effect of the small sample size used in this study. The sample size again may not allow the findings of this study to be extrapolated beyond KIU-Dar.

Hence, the research hypothesis suggested at the beginning of this study (cf. section 1.11) that the psychometric properties in resource challenged institutions could be poor compared to their counterparts in the developed countries may likely to be true, however, in addition to the limited resources, other factors such as lack of adequate training, low motivation could also have contributed to the poor quality of the OSCEs conducted in this study. Further work is required to be carried out in these resource limited schools to clarify this assumption.

6.2.5 Research question 5

The research question was stated as:

How can the findings of a psychometric analysis and interpretation be used to improve the OSCE at KIU-Dar over a period of time?

For research Question 5, the following objective was pursued:

Describe how the findings of the analysis and its interpretation can be used to improve the OSCE at KIU-Dar in Dar es Salaam over a period of time.

In Table 6.1 below, showing the station analysis report, using the difficulty index, discrimination index and the statistical significance (cf. section 5.2.3, cf. Table 5.7), the stations with good properties should be stored safely in an OSCE bank for future use, the ones with bad properties should be discarded and the ones with insufficiently good properties should be reviewed before they can be used in the future. The discarded stations should be replaced with better ones. Moreover, the stations with negative correlations and very poor reliability estimates should also be discarded (cf. section 5.2.4.1, cf. Tables 5.8 & 5.9).

TABLE 6.1 STATIONS ANALYSIS REPORT

ACTION	KEEP	REVIEW	DISCARD
Subject	*Good*	*Not Sufficient*	*Bad*
OBGYN	----	2&4	1&3
%		2(50%)	2(50%)
Paediatrics	1,4,5,12,13,20,21 & 26	2,6-11,15,17,18,22-25	3,14,16,19 & 27
%	8(29.6%)	14(51.9%)	5(18.5%)
Medicine	8 &18	1-4,7,9-17,19-24	5,6&25
%	2(7.7%)	21(80.8%)	3(11.5%)
Surgery	2,5,8-13,15-20	3,4,6,7,14,21-25,23	1
%	14(56%)	10(40%)	1(4%)

The discussions and recommendations provided in Chapter 5 and section 6.3 below further addressed this objective.

6.3 LIMITATIONS OF THE STUDY

The following challenges were encountered in the implementation of this study:

Even though the objectives of the study were very clear, it became a very comprehensive study because of the huge amount of data it generated. The study was conducted in the field of Health Professions Education focussing on quantitative analysis of the quality of the OSCE. However, some important areas of test quality such as the validity (qualitative), estimating the cost of the OSCEs, feasibility and educational impact were also mentioned and thereby broadening the scope of this study. These aspects were discussed briefly, but could be addressed in more detail when publications are prepared or new research topics are pursued.

The study was conducted especially for the OSCE, but it could also be relevant for other types of assessments and in other faculties and universities to improve the quality of examinations in the University as a whole. These aspects could be addressed in further research projects and studies.

The venue of the examination was not very conducive for the OSCE as the hospital is basically providing health care and not training. The space available for the OSCEs was not adequate as free movement of students and examiners were restricted. Staff, care-givers and visitors' activities during the OSCE may have contributed to the unsuitability of the venue. It is more convenient to conduct examinations in the university's own teaching hospital, because the clinical departments will be at liberty to implement the OSCEs at the time and in the manner they deem it fit to be carried out. This will be our experience as soon as the teaching hospital is completed. These factors, though may be part of the external errors influencing the scores of the candidates, they however, did not significantly affect the process of observation nor the collection of data during the OSCE.

Access to the OSCE results was challenging in terms of promptness in submission of results by the heads of departments to the researcher. Hence the researcher could not do a detailed analysis of the checklist itself. If it were possible in the future studies, the researcher should be permitted to access the scores and the completed examiners' checklist directly from the examinations office.

The unfamiliarity and lack of knowledge of the examiners in the area of global scoring limited the analysis of the scoring in this research. The absence of simulated patients also limited the richness of the findings of this research. The marking of the written stations were co-marked, hence, one examiner was not solely responsible to mark one full station. Therefore, it was difficult to match the examiners with each station and therefore clearly identify 'Hawks' and 'Doves' for the purposes of retraining and effective post examination remediation. Moreover, since the examinations are scheduled by the university, the research can only be carried out during examination seasons.

The small sample size (less than 30 students) available for this research would require some caution to be exercised in the extrapolation of these findings as a reflection of the OSCE practice in other Sub-Saharan Universities. Hence, further studies with larger sample size are necessary for a better representation of the OSCE practice in other universities.

Though the overall psychometrics was better than the individual subject metrics in this study, this cannot serve as an indicator of the quality of the examination process since each student need to pass each individual course to graduate.

Obviously, the investigation of the qualities of the OSCEs implementation in resource constrained institutions was not exhaustive in this study, further research need to be done and the scope broadened to include other types of examinations, other examination seasons, other faculties and similar resource constrained universities to give the medical education world a clear view of the OSCEs implemented in resource constrained universities, so that interventional efforts may be put in place to improve the OSCEs and therefore the quality of our medical graduates and patient care in this part of the world.

6.4 RECOMMENDATIONS, CONTRIBUTION AND SIGNIFICANCE OF THE RESEARCH

6.4.1 Recommendations

In order for the readers and institutions to benefit maximally from the content of this research, the researcher made the following recommendations:

6.4.1.1 *Specific recommendations for KIU-Dar*

i. OSCEs should be conducted in appropriate Hospitals/venues constructed and functioning for the purpose of training and clinical assessment mainly.
ii. To use other methods to test clinical skills assessments in addition to OSCE.
iii. Blueprinting the OSCEs is absolutely necessary.
iv. Introducing SPs into the OSCEs to improve the reliability of the OSCEs.
v. Design checklist to contain global scores.
vi. Use z-scores for the university grading system.
vii. Use borderline methods to determine cut-off scores.
viii. Station number should range between 8-20.
ix. The length of the stations should match the objectives and the tasks of each station.
x. Administering OSCEs in small sets of 2-4 cases over time and then pooling as a collective OSCE.
xi. In this study, all stations with low internal consistency coefficients should be completely revised or excluded, moderate coefficients should be reviewed for further improvement, and stations with high coefficient should be stored in secure OSCE stations bank (cf. Chapter 5, cf. Tables 5.8 & 5.9).
xii. Examiners, who are consistently matched to low internal consistency coefficient stations should be scrutinised by retraining them on writing, rating checklists and marking or replaced.
xiii. The un-attempted station (station 23: Surgery) should be investigated, completely revised, the topics addressed comprehensively in the training of subsequent candidates and a compensation plan to cover the unfamiliar topics for the outgoing students should

be made before they are certified.

xiv. Better training and learning methods to be instituted for adequate and proper clinical exposure.

6.4.1.2 *General recommendations*

i. Psychometric analysis should be incorporated into the Quality Assurance Policy of the university to ensure quality in the examination processes.
ii. All students' scores should be normalised to Z-scores.
iii. Grading system in the university should be based on z-scores.
iv. Regularly apply psychometric analysis on the summative university examinations.
v. Intensive in-house training of academic staffs and examiners in psychometric analysis and the OSCEs.
vi. Conducting seminars, workshops and conferences involving other universities on Psychometric analysis and the OSCEs to obtain a uniform level of understanding and skills in administering psychometric analysis and the OSCEs in resource limited universities.
vii. Establishing a medical education department in every university to supervise medical training.
viii. Examiners need to work closely with psychometricians to objectively measure the quality of the OSCEs and improve its practice in medical schools.
ix. More research needs to be done to further unfold the psychometric properties of the OSCE in this university and other medical schools with limited resources.
x. Medical schools should purchase the novel psychometric analysis programme software tool (Tavakol & Doody 2016:104) that can rapidly analyse OSCE scores within 3 minutes after entering the marks.

Even though psychometric analysis has been practiced for over 40 years in the developed countries, it is a relatively new knowledge in resource limited schools, of the third world regions like East Africa. Hence, this research made a valuable contribution to 'new' knowledge in resource constraint institutions by providing recommendations for developing and implementing the guidelines for regular post-examination psychometric analysis of the university assessments with the goal of improving its qualities. By developing the recommendations and guidelines, the identified gaps of lack of awareness, knowledge and skills in psychometric analysis amongst medical teachers in third world countries were bridged. The research can assist examiners to objectively measure the quality of examinations and over time improve the properties of their examinations, improve the quality of medical graduates and eventually patient care. The sound research approach and methodology ensured the quality, reliability and validity of the research. The completed research can form the basis for a further research agenda.

Recommendations made from the research will significantly improve teaching, training and learning in medical schools. Training will be more patient-centred with more Teacher-student interactions taking place on the patient wards rather than in the classrooms. Psychometric analysis in conjunction with other components of the quality assurance system will greatly improve the quality of the OSCEs and other types of examinations in resource limited institutions and other universities. Psychometric methods are valuable tools in the hands of the examiners.

The overall goal of this study was to investigate the psychometric methods available, illustrate its use and measure the psychometric properties of the OSCEs in Kampala International University, Dar campus as an example of resource constrained institution in East Africa in order to facilitate wider uptake of the practice of psychometric analysis of assessment processes in neighbouring institutions. The study provided clear recommendations to reach the goal that was set. The recommendations in achieving this goal were discussed above (cf. Section 6.3)

6.4.1.3 *The guideline for performing Psychometric analysis on the OSCE*

This guideline for carrying out psychometric analysis on the OSCE is simple and user - friendly. The tools needed are simple but good statistical programme software readily available in resource limited institutions such as the Microsoft Excel or the SPSS. In the chart below are the recommended eighteen steps to psychometric analysis, Figure 6.1. Performing only one of these tests will not produce a comprehensive picture of the quality of the exam nor identify the sources of errors for the purpose of improving the stations and the examination subsequently. The battery approach, whereby multiple tests are applied on the scores is the gold standard. The departments need to work closely with the psychometrician to achieve the goal of objectively measuring the quality of and improving the OSCEs. This guideline can also be applied to other forms of examinations.

The following are the steps to conducting psychometric analysis of the OSCE:

- Before the analysis commences, the biodata of the examiners, students and patients are obtained.
- Thereafter, the OSCE blueprint is used to categorise the stations for meaningful interpretation of the results.
- Then, the station maximum marks are harmonised for easy analysis and
- Incomplete results are excluded from the analysis.
- The refined data are entered into the statistical programme such as Microsoft Excel, graph pad or SPSS.
- The data are then restructured into multivariate and univariate formats. Multivariate for descriptive statistics (A) and univariate for inferential statistics (B).
- In step AI, the distribution of the station scores are described.
- In step AII, the pass mark and other measures of centre are determined.
- Step AIII involves measuring the variations and errors in the data.
- In step BI, there is station analysis.
- In step BII, the examiners' checklist scores are analysed in detail and the relationship between the examiners' checklist and global scores is investigated.
- The patients' rating of the students' performance is analysed in step BIII.
- In step BIV, the reliability of the OSCE is estimated.
- In step BV, the 3 steps to identify hawk and dove examiners are carried out.
- In step BVI, other properties of the OSCE can be assessed such as differences in the candidate gender scores, correlation between examiners characteristics and G-studies.
- After the analysis, post-examination remediation is carried out and finally, the next examination is improved based on the findings of the current psychometric findings.

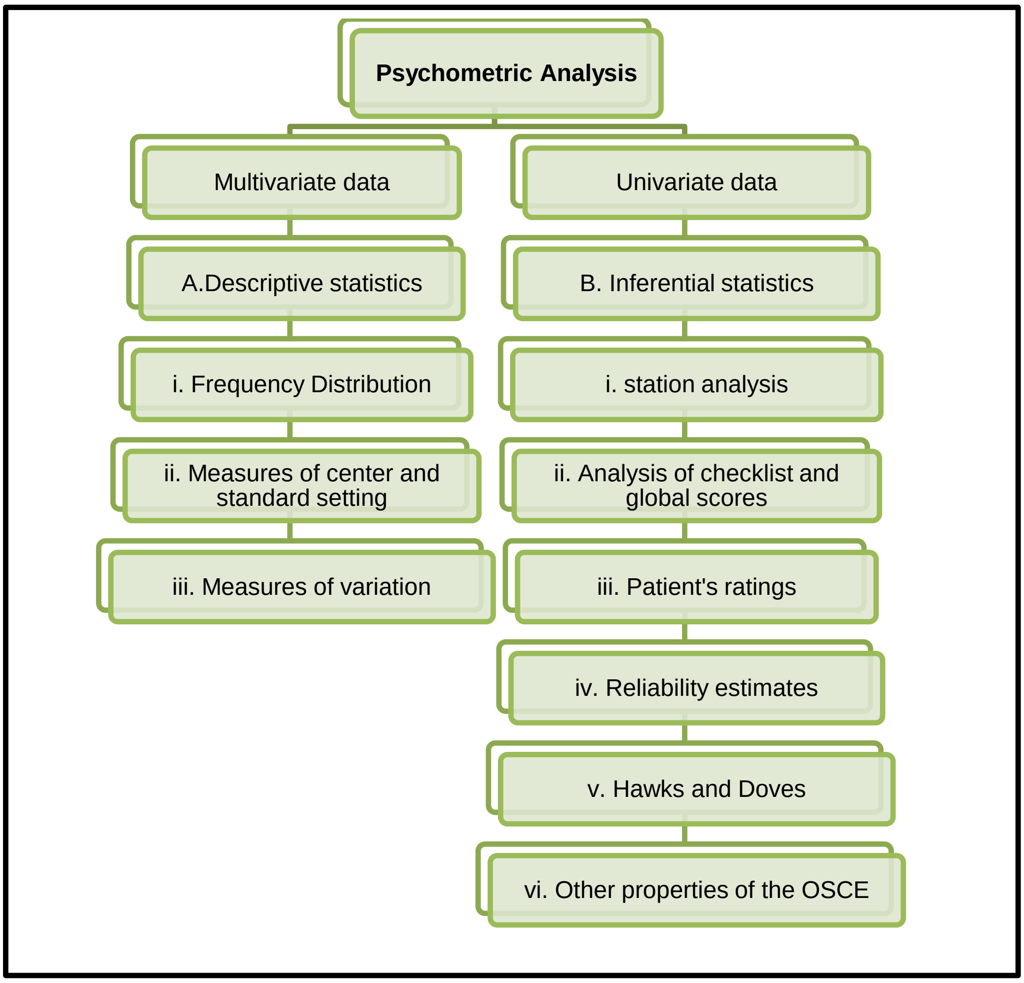
Psychometric Analysis
Multivariate data
Univariate data
A.Descriptive statistics
B. Inferential statistics
i. Frequency Distribution
i. station analysis
ii. Measures of center and standard setting
ii. Analysis of checklist and global scores
iii. Measures of variation
iii. Patient's ratings
iv. Reliability estimates
v. Hawks and Doves
vi. Other properties of the OSCE

FIGURE 6.1: GUIDELINE FOR PSYCHOMETRIC ANALYSIS
[COMPILED BY THE RESEARCHER, OGAH APRIL 2016]

The description of the application of the tests on the raw OSCE scores is cf. N in Chapter 2. Using this guideline which is manual, it took the researcher about one week to complete the analysis.

However, this year, Tavakol and Doody (2016:104) developed and published a rapid psychometric analysis programme software that can produce all the key quality statistics of an OSCE within 3 minutes after entering the scores into the programme. This new software may be out of the reach of the medical schools with resource challenges at the moment, therefore psychometric analysis can still be achieved manually with some adjustment to reduce the length of period spent on analysis.

6.5 POINTERS TO FURTHER RESEARCH

This study is the first of its kind in the East Africa region. It revealed the facts about the psychometric properties of the OSCEs practiced in one of the resource limited medical schools in the region. The findings of this study can only be generalised to the OSCEs practiced in the faculty of medicine, Kampala International University, Dar es Salaam. However, this study has opened the door for further investigations in the area of psychometrics to be carried out in this region. The several areas identified during this study that can be further investigated to strengthen the hypothesis that the OSCE practice in resource limited medical schools of Africa are not exactly the same as that practiced in resource established schools in other part of the world, include:

- The quality and relationship between examiners' checklist scores and global scores.
- Detailed analysis of each component of the examiners' checklist scores.
- Using standardised patients to increase the number of clinical stations and improve quality of the OSCE.
- Examine in detail the different aspect of validity (content, construct, face, predictive and concurrent validity) of the OSCE in resource limited institutions.
- Measure objectively the acceptability, feasibility, educational impact and cost of the OSCE in resource limited institutions.
- Examine the relationship or correlations between the theory and clinical assessments (concurrent validity).
- Carry out similar studies in other faculties and universities.
- Compare findings in these areas between medical schools in developed and developing countries.
- Repeat similar studies in the same schools to monitor progress in the quality of the OSCE and other examinations over time.
- Carry out D-studies to predetermine the properties of the OSCE and avoid errors in assessments.

Moreover, comparison studies of sequential years' assessments should be carried out in the future to measure the effects of the introduction of routine psychometrics to the examination quality assurance policy in the university.

6.6 OVERVIEW OF THE STUDY

A comprehensive study was carried out with a view to develop a guideline for psychometric analysis and recommend for its incorporation into the quality assurance examination policy of every medical school (beginning with the University of the Free State and Kampala International University medical schools) for regular use to consistently and objectively measure the quality of the OSCEs and other high stake university examinations with a view to improving and harmonizing the quality of our assessments, clinical training, quality of our medical graduates and consequently patient care.

Assessment is the heart of every training institution. Assessment drives learning. Through assessments, the amount of learning acquired by the trainee can be measured and the programmes can be evaluated. No assessment is ever similar to another one. OSCE being a form of clinical assessment which is resource intensive, its administration may be compromised in quality especially in resource limited institutions. The quality assurance machinery in place in most medical schools currently, to monitor the quality of assessment involves human raters which can be subjective, biased and inconsistent. Psychometric analysis, offers a stable, objective and cheap means of measuring and improving consistently the quality of the OSCEs and all other forms of assessments.

A lot has been documented and published about the psychometric qualities of the OSCEs practiced in the medical schools of the developed countries, but even there, it has not yet been incorporated into the University policies for regular use. Hence, every university has been producing medical graduates with different levels of competencies. Very little has been published about the real state of the OSCEs implemented in resource constrained medical schools in Sub-Saharan Africa. This might pose a risk to patient safety especially in countries that do not have a national qualifying examination to harmonise and regulate the quality of assessments and medical graduates certified and registered for practice.

Hence, psychometric analysis must be fully integrated into the quality assurance examination policy of every medical school, to harmonise and improve the quality of assessments, training, medical graduates and therefore patient care.

In this study, the psychometric methods were described and its use illustrated in a simple manner to measure, interpret and discuss the psychometric qualities of the exit OSCE implemented in a resource limited medical school in East Africa (Dar es Salaam). The medical school selected in this study is private-owned, 4 years old as of the time of study and does not have its own teaching hospital yet (the Teaching Hospital is currently under construction).

The medical school runs a staff-student exchange training programme, where staff and students at different levels of training from Dar campus can criss-cross with other students from its sister campuses in Uganda and Nairobi. The Dar campus clinical students rotate in nearby referral hospitals (structured and function basically for patient care) which are affiliated to the university. Hence, the venue of the exit OSCE was in one of the regional referral hospitals in Dar es Salaam.

The research methods comprised of literature reviews and observations using checklists during the OSCEs by the researcher and two assistants. The literature review provided a background for a conceptual framework and contextualized the problem against related theory and research. The examiners collected the scores (data) by means of checklists. A total of 27 graduating clinical students were examined.

The findings in this study were as follows: The exit OSCE was conducted in four clinical departments (OBGYN, Paediatrics, Internal Medicine and Surgery). Twenty-seven out of 30 registered final year clinical medicine students sat for the July 2015 final OSCE. They passed through 82 stations and were assessed by 20 examiners altogether. The examiners were a mixture of consultants, specialists and medical officers. The examiners were too few in paediatrics, surgery and internal medicine and the examiner: station ratio was 1:4 in these departments. The OSCE stations were too many in three of the departments but too few in OBGYN (cf. section 2.5.2). The clinical stations were manned by one examiner each and the rest of the stations (75%) were in written format. The examiners were not very conversant with the OSCE, they were not aware of global scoring, standard setting and psychometric analysis. Currently our medical school uses a fixed university pass mark of 50% and the grading system is based on the raw scores.

Only 52% of the students (the regular students) were sitting for the final OSCE for the first time, the rest had attempted the final OSCE in their respective disciplines previously but failed (cf. section 4.1). Sixty percent of the examiners were medical officers with first degree certificate (MBCHB), but with at least three years working experience. The medical officers were responsible for marking the written stations and also assisted the specialists and consultants to assess students in the manned stations (cf. section 4.4). The OSCEs, across the departments, were overly dominated by written stations rather than practical stations which is not in line with the objective of the OSCE. The departments resorted to this design because of the limited number of available examiners. Moreover, there was no blueprint for any of the OSCE stations. It was noted that none of the manned stations assessed the students in the central nervous system and neonatology. The central nervous system and the neonate are areas in the curriculum that need to be taught by highly qualified and experienced teachers which are scarce in the region of Africa. The scarcity of human resource is also noted in OBGYN, where all the academic staff was part-timers and this contributed to their inability to come early for the pre-OSCE briefing and generate sufficient number of OSCE stations. The absence of standardised patients in the OSCE due to insufficient time for recruitment and training was documented, however real patients were used for the manned stations. The candidates' OSCE scores were compiled and subjected to 27 psychometric tests which were grouped into 6 categories (cf. section 4:1) as follows: frequency distribution, measures of central tendencies, measures of variability, station analysis, reliability checks and identifying 'Hawks' and 'Doves'.

There was also no blueprint for the OSCE design. All the patients that were used were real, which limited the researcher from obtaining their ratings of the students performances as this was not permitted by the hospital administration. In addition, the researcher could not directly access the completed examiners' checklists, carry out a detailed analysis of the checklist scores and obtain the examiners' global scores. The Ministry of Health checklist that was used by the examiners did not capture global scoring nor standardised patients' ratings.

As a result it is practically difficult to have an ideal examination setting, the raw OSCE scores are likely to be abnormally distributed with possibly outliers as we found in this study. Hence, to have an accurate perception of the students' performances, be able to make valid pass/fail decision and make comparison with other assessments anywhere, it is better to convert the raw scores into z-scores and the university grading systems should be built on z-scores. As was also observed in the analysis, the letter grades based on the z-scores corresponded to different raw score ranges in each station and subject. Moreover, to make an accurate pass-fail judgement on students' performances, it is better to use the 'gold standard' method of setting the pass mark especially in a situation where the examination setting is not ideal. The borderline regression method is the recommended method, however, where there are no global scores, the borderline method, as was demonstrated in this study can be used to determine the pass mark for every examination.

The variability of the OSCE scores was generally high, which is not desirable for a criterion referenced test. Also, the variability in an ideal OSCE should only come from the different abilities of the students, however, in this study, most of the variability in the students' scores was contributed by the examiners and their interactions with the students as noted in the ANOVA and the G-studies.

The station analysis in this study showed that the internal medicine OSCE was the best. The other subjects had poor discriminating power and difficulty index for a criterion referenced assessment. The internal consistency and stability of the OSCEs in this resource constrained institution, based on Cronbach alpha, were low (below 0.25). Several stations were either marked too leniently or too strictly in the study.

These findings appear to be different from that experienced in established institutions. Several reasons could be postulated for the difference: Could it be due to the limited resources which include lack of adequate training, clinical exposure, poor motivation? Could it be that there were errors in the recording and transfer of scores from the checklists and scripts to the excel sheets that were submitted to the researcher from the various departments? If the difference is due to data entry errors, could this error affect all the 4 clinical departments? These later assumptions are difficult to investigate because the OSCE scores constitute only 10% of the entire subject scores for the academic year and therefore not published separately by the university examinations office. However, the likelihood of an error in recording or transfer of scores, could be seen in the pattern of the measures of central tendencies in some of the stations. Usually, in a positively skewed distribution, the mean should be higher than the median and the median should be higher than the mode. In a negatively skewed distribution, the mode is higher than the median and the median is higher than the mean (cf. Chapter 5, section 5.2.1.1). Using an example, OBGYN station 4 is positively skewed, however, the atypical order of the measures of central tendencies is displayed (mean=4.4, median=5, mode=6). Unfortunately, there are very few accessible publications of psychometrics from similar resource challenged medical schools to make comparison with.

As it stands at the moment, the quality of these OSCEs was substandard especially in OBGYN. It shows deficiencies more in clinical skills than in the knowledge of the subject, which may be an indication of how these graduating students will perform in clinical practice in the future. This could be a reflection of the poor staffing, funding, motivation, training and facility in this private school which receives its income only from tuition fees and few research grants as of now. The students who passed these final OSCEs may have to be followed up all through their internship and thereafter, to monitor their progress. The Tanzania Medical Council usually conducts a national certifying examination at the end of internship, before awarding full registration and this might inform the university of the competencies of their graduates in the job market. The rest of the candidates supplementing or repeating will benefit from the improved training and assessments as a result of this study. These research findings will be presented to the university management for their consideration and action to salvage the examination process in the school. As examination drives training and learning, more effort, time, resources and regular psychometrics need to be invested into subsequent OSCEs for better outcome. The current OSCEs may be used as a reference for comparison to monitor improvement in subsequent examinations. The university plans to create a medical education unit to coordinate and control quality in staff and student training and assessments in the university. Therefore in the very near future, training and assessment will be standardised in KIU-Dar.

The study originated from the recognition that a gap exists in the knowledge and skills of psychometric methods at KIU-Dar, the UFS, but also in Sub-Saharan Africa and the rest of Africa. Psychometrics involves statistics and mathematics, which presents a challenge to many practitioners. To bridge the gap, the researcher compiled a very simple guideline for psychometric analysis and recommendations for the integration of psychometric analysis of high stake examinations including the OSCEs as a required component in the current quality assurance examinations system in order to measure objectively and consistently the quality of examinations and improve the quality of high stakes examinations, training, learning, medical graduates and patient care.

The user-friendly psychometric guideline including the simple available tools and their applications were discussed with the goal to ensure that every examiner acquires the knowledge and skills of psychometric analysis and will be able to regularly apply it to their examinations in their institutions to improve their assessments. Moreover, a new psychometric programme for the rapid analysis of the OSCE data, developed and published by Tavakol and Doody (2016:104) was introduced in this study with recommendations for every Medical School to purchase it for use in their Medical Education Units to quality assure their examinations. By developing the strategy, the identified gap was bridged, in that it can aid in the training and encouragement of academic staff to use the psychometric tools skilfully on their subject examinations to improve assessments and training. Recommendations in this regard were made. The completed research can form the basis for a further research undertaking.

6.7 FINAL REMARKS

This study is the first of its kind in East Africa region. It revealed the facts about the psychometric properties of the OSCEs practiced in one of the resource limited medical schools in the region. In this study also, application of psychometric methods was illustrated and recommendations with guideline were produced for the regular use of psychometric analysis on the OSCE with the goal of improving assessments generally. The findings in this study support the assumption that the OSCE practice in resource limited schools may be of lower quality than that published in established schools. This assumption however needs further evidences. This study has opened the door for further studies in the area of psychometrics to be carried out in this region.

REFERENCES

Adamo, G. 2003. Simulated and standardized patients in OSCEs: achievements and challenges. *Medical Teacher; 25*(3):262–270.

Al-Naami, M.Y., El-Tinay, O.F., Khairy, G.A., Mofti, S.S., Anjum, M.N. 2011. Improvement of the psychometric properties of the Objective structured clinical examination when assessing problem solving skills of surgical clerkship. *Saudi Medical Journal*; 32(3):300-4.

Al-osail, M.A. Al-osail, E.M., Al-ghamdi, M.A., Al-hawas, A.M., Al-bahussain, A.S., Aldajani, A.A. 2014. Correlations between the Objective Structured Clinical Examination Score and Written Examinations in Undergraduates. *International Journal of Life Sciences Research*; 2(4):193-204.

Aranda, S., Yates, P. 2009. An overview of Assessment. *Canberra: The National Cancer Nursing Education Project (EdCaN), Cancer Nursing*. National Education Framework. Australia. Pp 2.
http://edcan.org.au/assets/edcan/files/docs/EdCan-Overview-of-Assessment.pdf
Retrieved on 13 October 2016.

Araoye, M.O. 2003. Research Methodology with statistics for Health and Social Sciences. Nathadex Publishers, Ilorin, First ed. Pp 55.

Avijiit, H., Nithya, G. 2016. IJD Module on Biostatistics and Research Methodology for the Dermatologist. *Indian Journal of Dermatology*; 61(1):10-20.

Bajwa, N.M., Belli,D., Vu, N.V., Park, Y.S. 2016. Improving the residency admissions process by integrating a professionalism assessment: a validity and feasibility study. Adv Health Sci Educ Theory Pract.
(http://www.ncbi.nlm.nih.gov/pubmed/27107883)
Retrieved on 24 June 2016.

Baker, L.M. 2006. Research Methods. Library Trends:55 (1):171–189.

Banerjee, A., Chaudhury, S. 2010. Statistics without tears: Populations and samples. *Industrial Psychiatry Journal*: 19 (1):60–65.

Banerjee, A., Chitnis, U.B., Jadhav, S.L., Bhawalkar, J.S., Chaudhury, S. 2009. Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*; 18 (2):127-31.

Barman, A. 2005. Critiques on the Objective structured Clinical Examination. *Annals of the Academy of Medicine, Singapore*; 34(8):478-82.

Bartman, I., Smee, S., Roy, M. 2013. Catching the Hawks and Doves: "A Method for Identifying Extreme Examiners on Objective Structured Clinical Examinations". *Clinical Teacher*; 10(1):27-31.

Biau, D.J., Jolles, B.M., Porcher, R. 2010. P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics Related Res*earch; 468 (3):885-92.

Blood, A.D., Park, Y.S., Lukas, R., Brorson, J.R. 2015. Neurology objective structured clinical examination reliability using generalizability theory. *Neurology*; 85(18):1623-9.

Borrego, M., Elliot P.D,C. 2009. Quantitative, Qualitative and Mixed Research Methods in Engineering Education. *Journal of Engineering Education;* 98(1):53-66.

Boulet, J.R., Mckinley, D.W., Whelan, G.P., *et al.* 2003. Quality Assurance Methods for Performance-Based Assessment. *Advances in Health Sciences Education*; 8(1):27-47.

Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smee, S., Sambandam, E. 2011. Performance in assessment: consensus statement and recommendations from the Ottawa conference. *Medical Teach*er; 33(5):370-83.

Boursicot, K.A.M., Roberts, T.E., Pell, G. 2007. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education;* 41(11):1024–1031.

Boursicot, K.A.M., Roberts, T.E. 2005. How to set up an OSCE. The *Clinical Teacher*; 2(1):16 – 20.

Bowen, G.A. 2005. Preparing a Qualitative Research-Based Dissertation: Lessons Learned. *The Qualitative Report*; 10(2):208-222.
http://www.nova.edu/ssss/QR/QR10-2/bowen.pdf
Retrieved on 31 May 2015.

Brannick, M.T., Erol-Korkmaz, H.T., Prewett, M. 2011. A systematic review of the reliability of objective structured clinical examination scores. *Medical Education;* 45(12):1181-9.

Breakwell, G.M., Hammon, S., Fife-Schaw, C. (Eds). 2000. *Research methods in psychology*. London: Sage Publications.

Bussiness Dictionary. Limited Resources.*Bussiness Dictionary.com*
http://www.businessdictionary.com/definition/limited-resources.html#ixzz36mV5e3sV
Retrieved on 16 September 2014.

Castleman, E.M. 2007. A model to manage continuous professional development for the alumni of a private higher education institution. (Unpublished PhD Thesis). University of the Free State.

Cook, D.A., Beckman, T., Bordage, G. 2007. Quality of reporting of experimental studies in medical education: A systematic review. *Medical Education;* 41(8):737-45·

Creswell, J.W., 2014. *Research Design: Qualitative, Quantitative and Mixed methods approaches*. Fourth edition. London; Sage Publications. pp:1-342.

Cronin, P., Ryan, F., Coughlan, M. 2008. Undertaking a literature review: A step-by-step approach. *British Journal of Nursing*: 17(1):38-43.

Crossley, J., Humphris, G., Jolly, B. 2002a. Assessing health professionals. *Medical Education;* 36(9):800-4.

Crossley, J., Davies, H., Humphris, G., Jolly, B. 2002b. Generalizability: a key to unlock professional assessment. *Medical Education;* 36:972–978.

Cruess, R.L., Cruess, S.R., Steinert, Y. 2016. Amending Miller's pyramid to include professional identity formation. *Academy of Medicine*; 91(2):180-5.

Downing, S.M. 2005. Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education*; 39:350-355.

Downing, S.M. 2004. Reliability: on the reproducibility of assessment data. *Medical Education;* 38:1006–1012.

Fan, A.P., Tran, D.T., Kosik, R. O., Mandell, G. A., Hsu, H. S., Chen, Y.S. 2012. Medical education in Vietnam. *Medical Teacher*; 34:103–107.
file:///C:/Users/HEALTH%20SCIENCE/Downloads/Medical%20education%20in%20Vietnam-hinari.pdf
Retrieved on 13 October 2016.

Ganesh, N.D., Sachin, K.H., Mohini, S.M. 2012. Basic Biostatistics for Postgraduate students. *Indian Journal of Parmacology;* 44(4):435-442.

Garger, J. 2010. 4 Levels of Measurement in Social Science Research. *John Garger.* https://johngarger.com/articles/methodology/4-levels-of-measurement-in-social-science-research
Retrieved on 8 June 2016.

George. D, Mallery, M. SPSS for Windows, step by step. A simple guide and reference. 17.0 update (10a edition). Boston: Pearsson.

Gupta P, Dewan, P, Singh, T. 2010. Objective Structured Clinical Examination (OSCE) Revisited. *Indian Paediatrics*; 47(11):911-20.

Hae, Y.K. 2014. Analysis of Variance (ANOVA) comparing means of more than two groups. *Restorative Dentistry and Endodontics*; 39(1):74-77.

Halder, A.K, Molyneaux, J.W, Luby, S.P., Ram, P.K. 2013. Impact of duration of structured observations on measurement of handwashing behavior at critical times. BioMed Central Open Access Publisher.
http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-13-705
Retrieved on 8 June 2016.

Hamza, M.A., Gominda, P., Zubair, A. 2014. A Comprehensive, Multi-modal Evaluation of the Assessment System of an Undergraduate Research Methodology Course: Translating Theory into Practice. *Pakistan Journal of Medical Sciences Online*: 30(2):227-32.

Harasym, P.H., Woloschuck, W., Cunning, L. 2008. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advances in Health Sciences Education*; 13(5):617-632.

Harden, R.M. 1988. What is an OSCE? *Medical Teacher*; 10(1):19-22.
Hemphill, J.F. 2003. Interpreting the Magnitude of Correlation Coefficients. *American Psychologist*: 58(1):78-9·

Heale, R., Twycross, A. 2015. Validity and reliability in quantitative studies. *Evidence- Based Nurs*ing; 18(3):66-7.

Hejri, S.M., Jalili, M., Muijtjens, A.M.M., Van Der Vleuten, C.P.M. 2013. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *Journal of Research in Medical Sciences*; 18(10): 887–891.

Institute for work and Health. 2015. What researchers mean by... cross-sectional vs. longitudinal studies. *Institute for work and Health Newsletter*; 81:1-2

Jebakumar, A.Z., Manoj, G. 2012. An overview of Biostatistics. *Journal of pharmaceutical biology*; 2(2):63-79.

Jones and Bartlett Learning. 2007. Research Evidence. Jones and Bartlett, Publishers. http://www.jblearning.com/samples/0763751081/5014x_ch02_015_024.pdf
Retrieved on 6 May 2016.

Kamran, Z.K., Sankaranarayanan, R., Kathryn, G., Piyush, P. 2013a. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: An Historical and Theoretical Perspective. *Medical Teacher*; 35(9):e1437–e1446.

Kamran, Z.K., Gaunt, K., Sankaranarayanan, R., Piyush, P. 2013b. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration*.* *Medical Teacher*; 35(9):e1447-e1463.

Kawulich, B. 2012. Collecting data through observation. In Doing Social Research: A global context, edited by C. Wagner, B. Kawulich, M. Garner. West Georgia: McGraw Hill, pp.150-160

Kim, H.Y. 2013. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*; 38(1):52–54.

Kim, T.K. 2015**.** T test as a parametric statistic. *Korean Journal of Anesthesiology;* 68(6):540-6.

Kim, J., Neilipovitz, D., Cardinal, P., Chiu, M. 2009. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). *Simulation in Healthcare*; 4(1):6-16.

Kimberlin, C.L., Winterstein, A.G. 2009. Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy*: 65(23):2276-84.

Kobayashi, K., Sakuratani, Y., Abe, T., Yamazaki, K., Nishikawa, S., Yamada, J., Hirose, A., Kamata, E., Hayashi, M. 2011. Influence of coefficient of variation in determining significant difference of quantitative values obtained from 28-day repeated-dose toxicity studies in rats. *Journal of Toxicological Sciences*; 36(1):63-71.

Krupski, T.L., Dahm, P., Fesperman, S.F., Schardt, C.M. 2008. How to perform a literature search. *The Journal of Urology*; 179(4):1264-70.

Kuzma, J.W., Bohnenblust, S.E. 2001. Basic Statistics for the Health Sciences, 4th eds. Mayfield Publishing Company; United States of America.

Labanca, F. (Ed.). 2010. Trustworthiness in Qualitative Research. *In search of Scientific Creativity, the weblog of Frank Labanca, Ed.D.*
http://problemfinding.labanca.net/2010/05/24/trustworthiness-in-qualitative-research/comment-page-1/
Retrieved/Downloaded on 16 September 2014.

Laerd Statistics. 2013. Cronbach's Alpha (α) using SPSS**.**
https://statistics.laerd.com/spss-tutorials/cronbachs-alpha-using-spss-statistics.php
Retrieved on 18 September 2014.

Laerd Statistics. 2013: Kurtosis in SPSS. 1 of 1.
https://statistics.laerd.com/spss-tutorials/kurtosis-using-spss-statistics.php
Retrieved on 18 September 2014.

Laerd Statistics. 2013. Measures of central tendencies using SPSS: 1 of 1.
https://statistics.laerd.com/spss-tutorials/measures-of-central-tendencies-using-spss-statistics.php
Retrieved on 18 September 2014.

Lennon-Dearing, R., Barnes, N. 2013. Quantitative Research. In H. R. Hall, L.A., Roussel (Eds). *Evidence-based practice: An integrative approach to research, administration, and practice*. Chapter: Quantitative Research. Memphis: Jones and Bartlett; pp.3-21.

Lewis, S.H., Wojcik, R. 2009. Methodologies for data collection. *BioMed Central, The Open Access Publisher*: 2(3):1-13.
Louise, A. 2002. Assessing Professional Behavior: Yesterday, Today and Tomorrow. *Academic Medicine;* 77(6):502–515.

Mandal, J., Acharya, S., Parija, S.C. 2011. Ethics in human research. Tropical Parasitology; 1(1):2–3.

Mangal, S.K., Mangal, S. 2013. Research Methodology in Behavioural Sciences. Delhi: PHI Learning Private Limited.
https://books.google.co.tz/books?id=uaVbAAAAQBAJ&pg=PA313&dq=data+collection+tools+in+research&hl=en&sa=X&ved=0ahUKEwi39eTcjubMAhWJcRQKHVD3AJYQ6AEIVDAG#v=onepage&q=data%20collection%20tools%20in%20research&f=false
Retrieved on 19 May 2016.

Mccrorie, P., Boursicot, K.A.M. 2009. Variations in medical school graduating examinations in the United Kingdom: Are clinical competence standards comparable? *Medical Teacher*; 31:223–229.

McDaniel, C., Gates, R. 2001. Marketing research essentials. 3rd edition. Ohio: South-Western College.

McLeod, S.A. 2015. Observation Methods. *Simply Psychology.*
www.simplypsychology.org/observation.html Retrieved on 18 May 2016. Pg 1-7.

McManus, I.C., Thompson, M., Mollon, J. 2006. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES): using multi-facet Rasch modeling. *BMC Medical Education;*
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1569374/
Retrieved on 24 June 2016.

McManus, I.C, Elder, A.T, De Champlain A, Dacre, J.E, Mollon, J., Chis, L. 2008. Graduates of different UK medical schools show substantial differences in performance on MRCP (UK) Part 1, Part 2 and PACES examinations. *BMC Medicine*; 6(1):5.

Miller, G.E. 1990. The assessment of clinical skills/competence/performance. *Academic Medicine*; Sep; 65(9):S63-7.

Mohammed, Y.A. 2008. Reliability, validity and feasibility of the objective structured clinical examination in assessing clinical skills of final year surgical clerkship. *Saudi Medical Journal;* 29(12):1803-1807.

Mohsen, T., Reg, D. 2011. Post-examination analysis of objective tests 1. *Medical Teacher;* 33(6):447–458.

Mookherjee, S., Monash, B., Wentworth, K.L., Sharpe, B.A. 2015. Faculty Development for Hospitalists: Structured Peer Observation of Teaching. Journal of Hospital Medicine.
http://www.researchgate.net/publication/259826593_Faculty_development_for_hospitalist s _Structured_peer_observation_of_teaching. Retrieved on 9 June 2016.

Mushquash, C., O'Connor, B.P. 2006. SPSS, SAS and MATLAB programs for generalizability theory analyses. *Behavior Research Methods*; 38(3):542-547.

Nerurkar, R.P. 2008. Basics of statistics for postgraduates. *Indian Journal of Dermatology Venereology Leprology*; 74(6):691-5.

Nick, T.G. 2007. Descriptive statistics. *Methods in Molecular Biology* ; 404:33-52.

Norcini, J.J. 2003. Setting standards on educational tests. *Medical Education*; 37(5):464-9.

Norcini, J.J. 2005. Current perspectives in assessment: the assessment of performance at work. *Medical Education*; 39(9):880-9.

Norcini, J., Anderson,B., Bollela, V., Burch, V., Joa, M., Costa, O., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V. A, Roberts, T. 2011. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*; 33(3):206-14.

Northern Arizona University. 2001. Module 2-Methods of Data Collection-Chapter 2-Online lesson. *Leisure Research Methods.*
http://www.prm.nau.edu/prm447/methods_of_data_collection_lesson.htm
Retrieved on 19 May 2016:1-14.

Nunnally, J.C., Bernstein, I.H. (Eds). 1994. *Psychometric Study*. New York: McGraw Hill, Inc.

Otwombe, K.N., Petzold, M., Martinson, N., Chirwa, T. 2014. A review of the study designs and statistical methods used in the determination of predictors of all-cause mortality in HIV-infected cohorts. *Pubmed*; 3;9(2):e87356.

Pell, G., Fuller, R., Homer, M., Roberts, T. 2010. How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49. *Medical Teacher;* 32(10):802–811.

Pérez-Vicente, S., Expósito, R.M. 2009. Descriptive statistics*.* *Allergologia et Immunopathologia*;37(6):314-20.

Pugh, D., Hamstra, S.J., Wood,T.J., Bordage, G. 2014. A procedural skills OSCE: assessing technical and non-technical skills of Internal Medicine Residents. *Advances in Health Sciences Education*; 20(1):1-16.

Putka, D.J., McCloy, R.A. 2008. Estimating Variance Components in SPSS and SAS: An Annotated Reference Guide 1. Human Resources Research Organization; 2 of 24.

Qualls, M.L., Pallin, D.J., Schuur, J.D. 2014. Parametric Versus Nonparametric Statistical Tests: The Length of Stay Example. *Academic Emergency Medicine.* 17(10):1113-21.

Raymond, M.R., Viswesvaran, C. 1993. Least squares models to correct for rater effects in performance assessment. *Journal of Educational Measurement*; 30(3):253-268.

Reinert, A. 2013. Assessment in Medical Education: A primer on methodology. http://ps.columbia.edu/education/sites/default/files/student_life/Assessment%20in%20Medical%20Education%20%20A%20Primer%20on%20Methodology%5B1%5D%20copy.pdf. Pg 25-34.
Retrieved on 23 June 2016.

Roberts, C., Newble, D., Jolly, B., Reed, M., Hampton, K. 2006. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical teacher;* 28(6):535–543.

Roberts, C., Rothnie, I., Zoanetti, N. *et al*. 2010. Should candidate scores be adjusted for interviewer stringency or leniency in multiple mini-interviews? *Medical Education;* 44(7):690–698.

Rozycki, W. 2003. Practicing Medicine. The Art and Science of Medicine. Volume XXVI (1).

Salkind, N.J. 2000. Exploring Research. 4th edition. Upper Saddle River, New Jersey: Prentice Hall.

Sedgwick, P., Greenwood, N. 2015. Understanding the Hawthorne effect. British Medical Journal (online).
https://www.researchgate.net/publication/281461000_Understanding_the_Hawthorne_effect.
Retrieved on 8 June 2016.

Schisterman, E.F., Whitcomb, B.W., Mumford, S.L., Platt, R.W. 2009. Z-scores and the birthweight paradox. *Paediatric and Perinatal Epidemiology*; 23(5):403-13.

Smith, J., Noble, H. 2016. Reviewing the literature. *Evidence Based Nursing*; 19(1):2-3.

Schwartz, A. 2011. Assessment in Graduate Medical Education: *A Primer for Pediatric Program Directors.* Chapel Hill, NC: *American Board of Pediatrics*; 1-121.

Steven, M. 2011. Objective Structured Clinical Examinations (OSCEs), psychiatry and the Clinical assessment of Skills and Competencies (CASC) Same Evidence, Different Judgement. *BMC Psychiatry*; 11(1):85.

Sullivan, A., Steven, M.S. 2003. Developing Countires*. Wikipedia, The Free Encyclopedia* (http://en.wikipedia.org/wiki/Developing_country)
Retrieved/Downloaded on 16 September 2014.

Taylor, C. 2014. How Do We Determine What Is an Outlier? *About education*.
http://statistics.about.com/od/Descriptive-Statistics/a/How-Do-We-Determine-What-Is-An-Outlier.htm
Retrieved on 24 June 2016.

Tavakol, M., Dennick, R. 2011a. Making sense of Cronbach's alpha. *International Journal of Medical Education;* 2:53-55.

Tavakol, M., Dennick, R. 2011b. Post-examination analysis of objective tests1. *Medical Teacher;* 33(6):447-458.

Tavakol, M., Dennick, R. 2012. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Medical Teachers;* 34(3):e161-e175.

Tavakol, M., Dennick, R. 2013. Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teachers;* 35(1):e838-e848.

Tavakol, M., Doody, G.A. 2015. Making students' marks fair: standard setting, assessment items and post hoc item analysis. *International Journal of Medical Education*; 6:38-39.

Tavakol, M., Doody, G.A. 2016. A novel psychometric programme for the rapid analysis of OSCE data: *Medical Teacher;* 38 (1):104-105*.*

The Web Centre for Social Research Methods. 2004. The t-test.
<http://www.socialresearchmethods.net/kb/stat_t.htm>
Retrieved on 23 May 2016.

Trochim, W.M.K. 2008. Sampling. Research Methods Knowledge Base. (http://socialresearchmethods.net/kb/sampling.php) Retrieved on 20 May 2016.
Turner, J.L., Dankoski, M.E. 2008. Objective Structured Clinical Exams: A Critical Review. *Family Medicine*; 40(8):574-8.

Thompson, C.B., Panacek, E.A. 2007. Research study designs: Non-experimental. *Air Medical Journal:* 26(1):18–22.

Thorndike, E.L. 1920. A constant error on psychological rating. *Journal of Applied Psychology IV*; 25-29.

Umar, A., Folorunsho, O. 2012. Sampling and Measurements in Course Guide:Data Collection Methodology,National Open University of Nigeria, School of Science and Technology. 1st ed. Pg 64-65.

Van der Vleuten, C. 2000. Validity of final examinations in undergraduate medical training. *Education and debate British Medical Journal;* 321:1217-1219.

Van der Vleuten, C., Schuwirth, L.W. 2005. Assessing professional competence: from methods to programmes. *Medical Education;* 39(3):309-17*.*

Verdaasdonk, E.G.G., Stassen, L., Widhiasmara, P.P., Dankelman, J. 2009. Requirements for the design and implementation of checklists for surgical processes. *Surgical Endoscopy*: 23(4):715-26.

Wallenstein, J., Ander, D. 2015. Objective structured clinical examinations provide valid clinical skills assessment in emergency medicine education. *The Western Journal of Emergency Medicine*; 16(1):121-6.

Wass, V. 2005. Ensuring medical students are "fit for purpose". *British Medical Journal (online*); 331(7520):791–792.

Williams, C. 2007. Research Methods. *Journal of Business and Economic Research*; 5(3):pp:65-72.

Yusoff, S., Bee, W.Y. 2012. Comparison of conventional measures of skewness and kurtosis for small sample size. (Paper presented at the Conference: Statistics in Science, Business, and Engineering (ICSSBE), International Conference.

Zygmunt, L.W., Marian, K. 2015. Statistical Properties of Skewness and Kurtosis of Small Samples from Normal and Two Other Populations. *Research Gate.*
https://www.researchgate.net/publication/286467970_Statistical_Properties_of_Skewness_and_Kurtosis_of_Small_Samples_from_Normal_and_Two_Other_Populations
Retrieved on 31 May 2016.

APPENDICES

Appendices attached to this thesis include:

Appendix A: Letter to the Dean of the School of Health Sciences to request permission to execute the study.

Appendix B: Letter to the Deputy Vice Chancellor-Academics

Appendix C: Letter to the participants requesting for their participation in the examination

Appendix D: Consent form

Appendix E: Research Instruments

APPENDIX A

LETTER TO THE DEAN SCHOOL OF HEALTH SCIENCES, KAMPALA INTERNATIONAL UNIVERSITY CONSTITUENT COLLEGE, DAR ES SALAAM, TANZANIA

15/07/2014

To: The Dean
School of Health Sciences
Kampala International University Constituent College
Dar es Salaam, Tanzania

Dear Sir,

Requesting permission to conduct PHD research

I humbly request for your permission to carry out my PHD research in the clinical departments of the School of Health Sciences, Kampala International University Constituent College, Dar es Salaam. The departments are: Paediatrics, Obstetrics & Gynaecology, Psychiatry, Medicine and Surgery and the title of the research is:

'Psychometric Analysis as a Quality Assurance System in OSCEs in a Resource Limited Institution. 'I am currently in the first year of my PHD programme (part-time) in Health Professions Education of the University of the Free State, South Africa and would like to carry out this study in each clinical departments with the current Final year Medical Students during their exit OSCE (objective structured clinical examination) evaluation. The research would involve observing and describing the OSCEs and analyzing the post-OSCE scores.

I will be glad if my request is granted.

Thank you for your consideration in advance.

Yours sincerely,

Dr Ogah Adenike Oluwakemi

APPENDIX B

LETTER TO THE DEPUTY VICE CHANCELLOR-ACADEMICS, KAMPALA INTERNATIONAL UNIVERSITY CONSTITUENT COLLEGE, DAR ES SALAAM, TANZANIA

15/07/2014

To: The Deputy Vice Chancellor-Academics
Kampala International University Constituent College
Dar es Salaam, Tanzania

Dear Sir,

Requesting permission to conduct PHD research

I humbly request for your permission to carry out my PHD research in the clinical departments of the School of Health Sciences, Kampala International University Constituent College, Dar es Salaam. The departments are: Paediatrics, Obstetrics & Gynaecology, Psychiatry, Medicine and Surgery and the title of the research is:

'Psychometric Analysis as a Quality Assurance System in the OSCEs in a Resource Limited Institution. 'I am currently in the first year of my Ph.D. programme (part-time) in Health Professions Education of the University of the Free State, South Africa and would like to carry out this study in each clinical departments with the current Final year Medical (MBCHB) students during their exit OSCE (objective structured clinical examination) evaluation. The research would involve observing and describing the OSCEs and analysing the post-OSCE scores.

I will be glad if my request is granted.

Thank you for your consideration in advance.
Yours sincerely,

Dr Ogah Adenike Oluwakemi

APPENDIX C

LETTER TO THE PARTICIPANTS (SIMULATED PATIENTS), KAMPALA INTERNATIONAL UNIVERSITY CONSTITUENT COLLEGE, DAR ES SALAAM, TANZANIA

15/07/2014

To: The Participants (Simulated Patients)
Kampala International University Constituent College
Dar es Salaam, Tanzania

Dear Sir,

Requesting for participation in PHD research

I humbly request for your participation in the study titled:

‘Psychometric Analysis as a Quality Assurance System in the OSCEs in a Resource Limited Institution. ‘The study will be conducted with the current Final year Medical (MBCHB) students during their exit OSCE evaluation.

I am currently in the first year of my PHD programme (part-time) in Health Professions Education of the University of the Free State, South Africa. The research would involve observing and describing the OSCEs and analysing the post-OSCE scores. If you accept, your role will be to fill a one page consent form and questionnaire which will not take more than two minutes of your time.

I will be glad if my request is granted.

Thank you for your consideration in advance.
Yours sincerely,

Dr Ogah Adenike Oluwakemi

APPENDIX D

INFORMATION DOCUMENT

FOR EXAMINERS

Study title: **Psychometric Analysis as a Quality Assurance System in OSCE in a Resource Limited Institution**

Greeting:

Introduction:

We, Dr Ogah Adenike Oluwakemi and research team, are doing research on Psychometric Analysis as a Quality Assurance System in OSCE in a Resource Limited Institution.

Research is just the process to learn the answer to a question. In this study we want to learn about the quality of the objective structured clinical examination (OSCE) that the students will be doing in the School of Health Sciences, Kampala International University Dar es Salaam Campus come February and July 2015.

Invitation to participate: We are asking you to participate in this research study as an examiner*.*

What is involved in the study – The study design is descriptive, where we want to describe and measure objectively the quality of the OSCE. The OSCE is simply a clinical form of examination that students of health sciences do before they can qualify to graduate or to be promoted to the next class. OSCE involves several stations with different tasks for the students to perform and they will be evaluated based on their performances in each station. You will be occupying a station and assessing every student that comes to your station. A brief examiners' training will take place a week before the examination day, conducted by the respective clinical department. On the day of the examination, you will again be briefed before the examination commences by the Head of the department. The researcher and her team will be observing the OSCE proceedings without obstructing any part of the examination. The data will be the students' scores that will be provided by the examiners and the opinion of the simulated patients at the end of this examination. This OSCE will last for 3hours. About 73 students from KIU will be involved in the examination.

Cost: You will not incur any monetary cost in this study.

Risks: There are no risks involved in this study.

Benefits: Refreshments during the examination will be provided by the school for your participation.

You will be given pertinent information on the study while involved in the project and after the results are available. The outcome of the study will be supplied to you at completion. The study will be presented at conferences and also published in peer-review journals.

Participation is voluntary, and refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled. You may discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled.

Reimbursements: You shall be reimbursed for any "out of pocket" expenses related to the study which has not been catered for as soon as enough evidences are provided.

Confidentiality: Efforts will be made to keep personal information confidential. Absolute confidentiality cannot be guaranteed. Personal information may be disclosed if required by law.
Organizations that may inspect and/or copy your research records for quality assurance and data analysis include groups such as the Ethics Committee for Medical Research and the Medicines Control Council.
If results are published, this may lead to individual/cohort identification.

Feedback: You will receive a copy of the study report at completion.

Publication: The study report will be presented in seminars and conferences and published in peer-review journals at completion.

Contact details of researcher(s) – for further information, please contact Dr Ogah Adenike Oluwakemi of the School of Health Sciences, Kampala International University Dar es Salaam Campus, Tanzania. Telephone number +255(0) 684934417.

If there are ethical queries, please contact the Secretariat and Chair: Ethics Committee of the Faculty of Health Sciences, University of the Free State – for reporting of complaints/problems: Telephone number +27(051) 4052812 OR Prof Rugamalira, Director of Postgraduate and Research, Kampala International University Dar es Salaam Campus, Tanzania. Telephone number +255(0) 713304032.

LETTER TO OBTAIN CONSENT FROM THE PARTICIPANTS

FORM EC 31

SAMPLE OF CONSENT DOCUMENT

(This document must be written in a language understandable to the participant)

CONSENT TO PARTICIPATE IN RESEARCH

FOR EXAMINERS

PROJECT TITLE: **Psychometric Analysis as a Quality Assurance System in OSCE in a Resource Limited Institution**

You have been asked to participate in a research study.

You have been informed about the study by …Dr Ogah Adenike Oluwakemi and request that you permit the researcher to observe you during the examination and to access all your scores for this OSCE at the end of the examination.

You may contact Dr Ogah Adenike Oluwakemi at the School of Health Sciences, Kampala International University, Dar es Salaam Campus any time if you have questions about the research or if you are injured as a result of the research.

You may contact the Secretariat of the Ethics Committee of the Faculty of Health Sciences, UFS at telephone number +27(051) 4052812 OR Prof Rugamalira, Director of Postgraduate and Research, Kampala International University Dar es Salaam Campus, Tanzania. Telephone number +255(0) 713304032, if you have questions about your rights as a research subject.
Your participation in this research is voluntary, and you will not be penalized or lose benefits if you refuse to participate or decide to terminate participation.

If you agree to participate, you will be given a signed copy of this document as well as the participant information sheet, which is a written summary of the research. The research study, including the above information has been verbally described to me. I understand what my involvement in the study means and I voluntarily agree to participate.

________________________ ___________________

Name and Signature of Participant Date

______________________ ___________________

Name and Signature of Witness Date *(Where applicable)*

______________________ ___________________

Name and Signature of Researcher Date

INFORMATION DOCUMENT

FOR STUDENTS

Study title: **Psychometric Analysis as a Quality Assurance System in OSCE in a Resource Limited Institution**

Greeting:

Introduction:

We, Dr Ogah Adenike Oluwakemi and research team, are doing research on Psychometric Analysis as a Quality Assurance System in OSCE in a Resource Limited Institution.

Research is just the process to learn the answer to a question. In this study we want to learn about the quality of the objective structured clinical examination (OSCE) that the students will be doing in the School of Health Sciences, Kampala International University Dar es Salaam Campus come February and July 2015.

Invitation to participate: We are asking you to participate in this research study as a student*.*

What is involved in the study – The study design is descriptive, where we want to describe and measure objectively the quality of the OSCE. The OSCE is simply a clinical form of examination that students of health sciences do before they can qualify to graduate or to be promoted to the next class. OSCE involves several stations with different tasks for the students to perform and they will be evaluated based on their performances in each station. You will be moving from one station to another and carrying out the task in each station. A briefing will take place 30 minutes before the examination begins by the respective Head of clinical Department. The researcher and her team will be observing the OSCE proceedings without obstructing any part of the examination. The data will be the students' scores that will be provided by the examiners and the opinion of the simulated patients at the end of this examination. This OSCE will last for 3hours. About 73 students from KIU will be involved in the examination.

Cost: You will not incur any monetary cost in this study.

Risks: There are no risks involved in this study.

Benefits: Refreshments during the examination will be provided by the school for your participation.

You will be given pertinent information on the study while involved in the project and after the results are available. The outcome of the study will be supplied to you at completion. The study will be presented at conferences and also published in peer-review journals.

Participation is voluntary, and refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled. You may discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled.

Reimbursements: You shall be reimbursed for any "out of pocket" expenses related to the study which has not been catered for as soon as enough evidences are provided.

Confidentiality: Efforts will be made to keep personal information confidential. Absolute confidentiality cannot be guaranteed. Personal information may be disclosed if required by law.

Organizations that may inspect and/or copy your research records for quality assurance and data analysis include groups such as the Ethics Committee for Medical Research and the Medicines Control Council.

If results are published, this may lead to individual/cohort identification.

Feedback: You will receive a copy of the study report at completion.

Publication: The study report will be presented in seminars and conferences and published in peer-review journals at completion.

Contact details of researcher(s) – for further information, please contact Dr Ogah Adenike Oluwakemi of the School of Health Sciences, Kampala International University Dar es Salaam Campus, Tanzania. Telephone number +255(0) 684934417.

If there are ethical queries, please contact the Secretariat and Chair: Ethics Committee of the Faculty of Health Sciences, University of the Free State – for reporting of complaints/problems: Telephone number +27(051) 4052812 OR Prof Rugamalira, Director of Postgraduate and Research, Kampala International University Dar es Salaam Campus, Tanzania. Telephone number +255(0) 713304032.

LETTER TO OBTAIN CONSENT FROM THE PARTICIPANTS

FORM EC 31

SAMPLE OF CONSENT DOCUMENT

(This document must be written in a language understandable to the participant)

CONSENT TO PARTICIPATE IN RESEARCH

FOR STUDENTS

PROJECT TITLE: **Psychometric Analysis as a Quality Assurance System in OSCE in a Resource Limited Institution**

You have been asked to participate in a research study.

You have been informed about the study by …Dr Ogah Adenike Oluwakemi and request that you permit the researcher to observe you during the examination and to access all your scores for this OSCE at the end of the examination.

You may contact Dr Ogah Adenike Oluwakemi ….. at the School of Health Sciences, Kampala International University, Dar es Salaam Campus any time if you have questions about the research or if you are injured as a result of the research. You may contact the Secretariat of the Ethics Committee of the Faculty of Health Sciences, UFS at telephone number +27(051) 4052812 OR Prof Rugamalira, Director of Postgraduate and Research, Kampala International University Dar es Salaam Campus, Tanzania. Telephone number +255(0) 713304032, if you have questions about your rights as a research subject.

Your participation in this research is voluntary, and you will not be penalized or lose benefits if you refuse to participate or decide to terminate participation. If you agree to participate, you will be given a signed copy of this document as well as the participant information sheet, which is a written summary of the research. The research study, including the above information has been verbally described to me. I understand what my involvement in the study means and I voluntarily agree to participate.

__________________________ ____________________

Name and Signature of Participant Date

________________________ ____________________

Name and Signature of Witness Date *(Where applicable)*

________________________ ____________________

Name and Signature of Researcher Date

FOR SIMULATED PATIENTS

Kichwa cha utafiti: Psychometric Analysis as a Quality Assurance System in OSCE in a Resource Limited Institution

Salamu:
Utangulizi:

Sisi **Dk. Ogah Adenike Oluwakemi** pamoja na kundi la utafiti tunafanya utafiti juu ya Psychometric Analysis as a Quality Assurance System in OSCE in a resource limited institution.

Utafiti ni kitendo cha kujifunza majibu juu ya swali husika. Katika utafiti huu tunahitaji kujifunza juu ya **OSCE** ambao wanafunzi watakuwa wanafanya katika chuo kikuu cha Kampala Internationl tawi la Dar es salaam kati ya Februari na Julai 2015.

Ukaribisho wa kushiriki: Tunakuomba kushiriki katika utafiti huu kama mgonjwa.

Vitu vinavyohusika katika utafiti: Njia ya kufanya utafiti iliyochaguliwa ni maelezo, ambapo tunahitaji kupima na kuelezea kwa ufasaha kiwango cha **OSCE**. OSCE kwa kifupi ni fomu ya kupimia kliniki ambayo wanafunzi wa shule ya sayansi ya afya wanafanya kabla ya kuruhusiwa kuhitimu masomo yao au kuruhusiwa kuingia katika darasa la ngazi nyingine. **OSCE** inajumuisha vituo mbalimbali na kazi tofauti kwa mwanafunzi kuzifanya na hufanyiwa tathmini kulingana na utendaji katika vituo husika. Wewe utakuwa upo kwenye kituo kama mgonjwa na hali ya ugonjwa unayotakiwa kuionyesha utaelezwa wiki moja kabla ya upimaji. Siku ikifika utaelezwa kabla upimaji haujaanza na unategemewa kujaza dodoso fupi kuhusu maoni yako juu ya utendaji wa kila mwanafunzi katika kituo chako. Kila mwanafunzi atakapokuja katika kituo chako, watakuuliza maswali na kukupima. Utawapima wanafunzi kama mgonjwa ambaye anapokea huduma kutoka kwa daktari na kama unaridhika na namna wanafunzi wanavyoshughulika na tatizo lako. Maelezo yako ndiyo zitakuwa maksi za wanafunzi watakazopewa na mtahini na wewe kama mgonjwa mwishoni mwa utafiti huu. **OSCE** hii itaisha kwa miezi mitatu. Takribani wanafunzi 70 kutoka chuo kikuu cha Kampala International tawi la Dar es Salaam wanahusishwa kwenye upimaji.

Hatari: hakuna hatari yoyote inayoambatana na utafiti huu.

Faida: Usafiri kwenda- kurudi na chakula wakati wa upimaji vitatolewa na Shule ya Sayansi ya Afya kwa ushiriki wako.

Ushiriki ni hiyari na kukataa kushiriki hakutahusisha adhabu yoyote au upotevu wa chochote kama uliahidiwa. Unaweza kukatisha ushiriki wako wakati wowote bila adhabu au upotevu wa chochote kama ulichoahidiwa.

Fidia: kama utakuwa umetumia pesa yako yoyote au gharama yoyote utarudishiwa mara tu itakapothibitishwa.

Siri. Juhudi za makusudsi zitafanyika kuhakikisha siri ya mtu inatunzwa. Utoaji wa siri utaruhusiwa endapo utahitajika kisheria tu.
Taasisi zinazoweza kukagua/ au kurudufu data zako kwa ajili ya kuhakikisha ubora na kuzifanyia tathmini ni pamoja na kamati ya maadili kwa ajili ya utafiti wa kiafya na kamati ya udhibiti wa madawa.

Kama matokeo yakiwekwa kwenye mitandao yanaweza kupelekea mtu/ kundi kujulikana

Mawasiliano Zaidi kwa watafiti: kwa maelezo Zaidi tafadhali wasiliana na **DK. Ogah Adenike Oluwakemi** wa shule ya sayansi ya afya, Chuo kikuu cha Kampala tawi la Dar es Salaam, Tanzania. Simu namba **+ 255(0) 684934417**

FOMU YA KUSHIRIKI

FOR SIMULATED PATIENTS

MAKUBALIANO YA WAHUSIKA WA UTATIFI

KICHWA: **Psychometric Analysis as a Quality Assurance System in OSCE in a resource limited Institution**

Unaombwa kushiriki kwenye utafiti

Ulitaarifiwa kuhusu utafiti na **Dk. Ogah Adenike Oluwakemi**

Unaweza kuwasiliana na **Dk. Ogah Adenike oluwanike** wa shule ya sayansi ya afya chuo kikuu cha Kampala International tawi la Dar es salaam wakati wowote kama una swali kuhusu utafiti au kama una wasiwasi wowote kuhusu matokeo yatakayotokana na utafiti huu.

Unaweza kuwasiliana na Katibu wa kamati ya maadili wa shule ya sayansi ya afya kupitia simu namba **+ 27(051) 4052812** au **Profesa Rugemalila** kwa namba **+ 255(0)713304032** kama una swali lolote kuhusu utafiti huu.

Ushiriki wako kwenye utafiti huu ni hiyari yako na hakuna adhabu juu yako kwa kukataa au kukatisha ushiriki wako kwenye utafiti huu. Kama utakubali kushiriki kwenye utafiti huu utapewa fomu iliyosainiwa ya utafiti huu na fomu ya ushiriki ambayo imeandikwa mafupisho ya jinsi utafiti utakavyokuwa.

Utafiti na maelezo hapo juu nimeezwa kwa maneno ya mdomo. Nimeelewa nini maana ya ushiriki wangu kwenye utafiti na kwa hiyari yangu nakubali kushiriki katika utafiti.

---   --------------------------------
Jina/Sahihi ya mshiriki/mtoto   Tarehe

--   ---------------------------------
Sahihi ya shahidi   Tarehe

____________________________   ____________________
Jina/Sahihi ya Ulitaarifiwa   Tarehe

CHILD ASSENT FORM

You are being asked to take part in a research study being done by the **Researcher, Dr Ogah Adenike Oluwakemi** in the **Kampala International University, Dar es Salaam Constituent College, Tanzania**. In this study, we are interested to **measure the quality of the OSCE (Objective Structured Clinical Examination)**. We have asked your parent or caregiver whether it is OK for you to participate, but now we want to see if it is OK with you.

If you decide to take part in this study, you will be given questionnaires to fill out **your opinion of the student's performance in your station**. This will take about 10 minutes to do. All the information we collect will be kept secret and you don't have to share any of your answers in the questionnaire with anybody else. We will not use your name so everything will remain private.

By signing this you are showing that you understand what is going to be happening and have asked any questions you may have about the research. You can also ask questions later if you cannot think of them now. Signing this form does not mean that you have to finish the study- you can pull out from the study at any time without explaining why.

________________________________ ______________

Child's signature Date

KISWAHILI TRANSLATION OF THE CHILD ASSENT FORM

Fomu Ya Makubaliano Ya Mtoto

Nimekuuliza kuchukua sehemu ya uchunguzi Dr Ogah Adenike Oluwakemiin the **Kampala International University, Dar es Salaam Constituent College, Tanzania**.

Kwa Masomo haya nina kukumbusha kupira ubora. Nimewuomba wazazi wako kama wapo tayari kushiriki Lakini usanataka kuona kama upo tayari.

Kama umekubali kuchuku malomo tunakupa karatafi ya kuja za. Kujaza maoni yako ya maliwu dhurio aumatokeo yako ambayo itachukua dakika 10 kufanya.

Au taarifa nyingine tutakisanya na itakuwa siri na hautanuwaye sha intu maswali yako kwa mtu yoyote. Hatutatumia jina lako kila kitu kitabaki kuwa siri.

Kwa saini hii karatasi inaoaye sha kuwa umeonyesha kuwu kitugani kitaendelea na tutakuuli za maswali tinaanini kwa unayotari ta ku husu kchunguzi pia unawe za kuuliza inaswahili baade kama hufikiri kwa susa.

Ku saini hii fomu haimaanishi kinua umemaliza masomo yako unaweza kuacha masomo yako bila kiyieleza.

……………………………………… ……………………..

Saini ya motto Tarehe

APPENDIX E

RESEARCH INSTRUMENTS

I. ---- RESEARCHER'S CHECKLIST FOR OBSERVATION DURING OSCE

II. --- EXAMINER'S CHECKLIST

III. -- QUESTIONNAIRE FOR SIMULATED PATIENTS

II. THE EXAMINER'S CHECKLISTS (4) FOR OBGY, PAEDIATRICS, INTERNAL MEDICINE AND SURGERY ARE THE PROPERTIES OF THE MINISTRY OF HEALTH, TANZANIA. SEE ATTACHMENT.

III. QUESTIONNAIRE FOR SIMULATED PATIENTS

Serial number:……………………..

Station:……………………………..

Gender:……………………………..

Age:…………………………………….

Did the Candidate engage you in:

A. Conversation?: yes…… No…….

If yes, was the conversation cordial and professional?(please tick)

1. Strongly agree
2. Agree
3. Neutral
4. Disagree
5. Strongly disagree

B. Physical examination?: yes…… No…….

If yes, was the examination gentle?(please tick)

1. Strongly agree
2. Agree
3. Neutral
4. Disagree
5. Strongly disagree

C. Procedure?: yes…… No…….

If yes, was the activity gentle?(please tick)

1. Strongly agree
2. Agree
3. Neutral
4. Disagree
5. Strongly disagree

D. Would you prefer this Doctor to attend to you when you need health care again?

Yes………… No……………….

If No,

why?..

....

....

IV. STATISTICAL PACKAGES

1. Linear regression and the coefficient of determination R^2 in SPSS (David, 2012:2).
2. How to Calculate R2 in Excel.
3. Cronbach's Alpha (ɑ) using SPSS
4. Standard Error of the Mean from SPSS
5. Pearson's Correlation in SPSS
6. ANOVA in SPSS

a. One-way ANOVA in SPSS(Laerd, 2013: 1 of 6)
b. Generalizability Study using SPSS:

After entering the examination data into the SPSS, the data is restructured from a multivariate design to a univariate design, whereby the primary outcome variable of interest (i.e. ratings) appears in only one column. Univariate formats will result in data sets with multiple records (rows) per object of measurement (e.g. persons), whereas multivariate formats will typically have only one record per object of measurement (Putka & McCloy (2004:2 of 24). After restructuring, then analysis can commence.

In SPSS, the analysis of variance components is carried out as described by Tavakol & Dennick (2012:e172): 'To this end, from the data menu at the top of the screen in SPSS, one clicks on 'restructure' and follows the appropriate instructions. Then to obtain the variance components, the following steps are carried out: From the menus choose 'Analyse', 'General Linear Model', respectively. Then click on 'variance components'. Click on 'Score' and then click on the arrow to move 'Score' into the box marked 'dependent variable'. Click on student and examiner to move them into 'random factors'. After 'variance estimates' appears, click OK and the contribution of each source of variance to the result will be presented'.

The G-coefficient (p^2) is defined as the ratio of the student variance component (denoted 'Vs') to the sum of the student variance component and the residual variance (denoted 'Ve') divided by the number of examiners (k) and written as follows: p^2 = Vs/Vs + (Ve/k) for a single facet design. In a multi facet design, p^2 = Vs/Vs +(Vi/k+Ve/k+Vst/k+Vsp/k) Tavakol & Dennick (2012:e173).

RESEARCHER'S CHECKLIST FOR OBSERVATION DURING OSCE

i.-- Type of stations and number:

1. Manned
2. Written
3. History
4. Physical examination
5. Diagnosis
6. Interpretation of Laboratory results
7. Proceedural
8. Treatment
9. Communication/counselling skills.

ii.-- Systems covered:

1. Central Nervous System
2. Cardiovascular System
3. Respiratory system
4. Gastrointestinal ystem
5. Musculoskeletal sytem
6. Neonatal

iii.-- Design in Parallel stations are the same.

1. Strongly aggree
2. Aggree
3. Neutral
4. Dissaggree
5. Strongly dissaggree

iii.--Tools in the stations follow the requirements of the checklist design.

1. Strongly aggree
2. Aggree
3. Neutral
4. Dissaggree
5. Strongly dissaggree

iv.-- All Assessors arrived early:

Yes or No

v.-- Assessors were briefed: Yes or No

vi.-- Students were briefed: Yes or No

vii.-- Patients were briefed: Yes or No

viii.-- Behaviour of assessors are appropraite.

1. Strongly aggree
2. Aggree
3. Neutral
4. Dissaggree
5. Strongly dissaggree

ix. -- Behaviour of simulated patients are appropraite.

1. Strongly aggree
2. Aggree
3. Neutral
4. Dissaggree
5. Strongly dissaggree

x. -- Behaviour of real patients are appropraite.

1. Strongly aggree
2. Aggree
3. Neutral
4. Dissaggree
5. Strongly dissaggree

Department......................Date......

RATING SCALES DOCUMENTATION

THE UNITED REPUBLIC OF TANZANIA

MINISTRY OF HEALTH AND SOCIAL WELFARE

TECHNICIAN CERTIFICATE IN CLINICAL MEDICINE

RATING SCALE FOR CLINICAL EXAMINATION IN INTERNAL MEDICINE

NAME OF SCHOOL --

Candidate's examination Number: --

The examiner shall observe closely when the student is demonstrating skills during history taking and physical examination, attitude when interacting with the patient and skills during the whole period of doing clinical examination

S/N	Component	Located marks	Attained score	Total section score	Comments
1	ATTITUDE DURING EXAMINATION			**6**	
	Appearance, dressing code and preparation (equip)	2			
	Introduced him/herself to the patient and informed the patient about the aim the procedure	1			
	Ensured patient/client privacy and confidentiality	1			
	Obtains informed consent	1			
	Appropriate patient positioning; ensure safety and comfort	1			
2	HISTORY			**15**	
	Asked about the chief complaints and duration in chronological order	2			
	Amplification of the History of presenting illness well detailed and explore appropriate information about the illness,	8			
	Review of other systems	1			
	Past medical history	2			
	Family and social history	2			
3	PHYSICAL EXAMINATION			**25**	
	Perform general examination appropriately	4			
	Perform systemic examination appropriately				
	Points to be added in the affected system	3			
	Cardiovascular system	4			
	Respiratory system	4			
	Per abdomen	4			
	Nervous system	4			
	Interpret signs	2			
4	**Summary and Presentation skills**			**8**	
	summary	3			
	All information obtained from the history well presented	3			
	All information obtained from physical examination well presented	2			

	Deduction of marks if reported findings not asked or done during history taking or physical examination	- 2.5			
5	**DIAGNOSIS**			**10**	
	Provisional diagnosis (Narrates features in support and against)	5			
	Provides and defends differential diagnoses	5			
6	**INVESTIGATIONS**			**6**	
	Provides the appropriate list and rationale of appropriate investigations	6			
7	**PATIENT CARE PLAN**			**15**	
	Discusses the most suitable treatment option according to the diagnosis/es				
	Correct medicine(s)	6			
	Correct dose	3			
	Correct schedule	3			
	Provided the follow-up plan and prognosis	3			
8	**PREVENTION**			**5**	
	Outline various modes of prevention	5			
	DISCUSSION IN RELATION TO PATIENT	**10**		**10**	
	Deduction for dangerous mistake **(Maximum deduction should be 10 marks)**				
	TOTAL SCORE			**100**	

NB: If the student reports something which he/she did not perform, half of the allocated marks for that specified area should be deducted. Likewise, if the student failed to perform a certain procedure during the course of history taking and physical examination in Serial Number 1, 2 and 3, the student will score 0 mark in that specified area.

Examiner's general comments --

Examiner's Name --------------------------------------- Signature ------------------------ Date --------------------

UNITED REPUBLIC OF TANZANIA

MINISTRY OF HEALTH AND SOCIAL WELFARE

TECHNICIAN CERTIFICATE IN CLINICAL MEDICINE

RATING SCALE FOR CLINICAL EXAMINATION IN OBSTETRICS AND GYNAECOLOGY

NAME OF SCHOOL --

Candidate's examination Number: --

The examiner shall observe closely when the student is demonstrating skills during history taking and physical examination, attitude when interacting with the patient and skills during the whole period of doing clinical examination.

S/N	Component	Located marks	Attained score	Total section score	Comments
1	**ATTITUDE DURING EXAMINATION**			**6**	
	Appearance, dressing code and preparation (equip)	2			
	Introduced him/herself to the patient and informed the patient about the aim the procedure	1			
	Ensured patient/client privacy and confidentiality	1			
	Obtains informed consent	1			
	Appropriate patient positioning; ensure safety and comfort	1			
2	**HISTORY**			**15**	
	Asked about the chief complaints and duration in chronological order	2			
	Amplification of the History of presenting illness well detailed and explore appropriate information about the illness,	5			
	Review of other systems	1			
	Obstetric history/ANC	3			
	Gynaecological history	3			
	Family and social history	1			
3	**PHYSICAL EXAMINATION**			**25**	
	Perform general examination appropriately	4			
	Perform systemic examination appropriately				
	Cardiovascular system	4			
	Respiratory system	4			
	Per abdomen (plus pelvic examination)	8			
	Nervous system	3			
	Interpret signs	2			
4	**Summary and Presentation skills**			**8**	
	Summary	3			
	All information obtained from the history well presented	3			
	All information obtained from physical examination	2			

	well presented				
	Deduction of marks if reported findings not asked or done during history taking or physical examination	-2.5			
5	**DIAGNOSIS**			**10**	
	Provisional diagnosis (Narrates features in support and against)	5			
	Provides and defends differential diagnoses	5			
6	**INVESTIGATIONS**			**6**	
	Provides the appropriate list and rationale of appropriate investigations	6			
7	**PATIENT CARE PLAN**			**15**	
	Discusses the most suitable treatment option according to the diagnosis/es				
	Correct medicine(s)	6			
	Correct dose	3			
	Correct schedule	3			
	Provided the follow-up plan and prognosis	3			
8	**PREVENTION**			**5**	
	Outline various modes of prevention	5			
	DISCUSSION IN RALATION TO PATIENT	**10**		**10**	
	Deduction for dangerous mistake **(Maximum deduction should be 10 marks)**				
	TOTAL SCORE			**100**	

NB: If the student reports something which he/she did not perform, half of the allocated marks for that specified area should be deducted. Likewise, if the student failed to perform a certain procedure during the course of history taking and physical examination in Serial Number 1, 2 and 3, the student will score 0 mark in that specified area.

Examiner's general comments ---

Examiner's Name ---Signature ---------------------------- Date ----------------------

THE UNITED REPUBLIC OF TANZANIA

MINISTRY OF HEALTH AND SOCIAL WELFARE

TECHNICIAN CERTIFICATE IN CLINICAL MEDICINE

RATING SCALE FOR CLINICAL EXAMINATION IN PAEDIATRICS AND CHILD HEALTH

NAME OF SCHOOL --

Candidate's examination Number: --

The examiner shall observe closely when the student is demonstrating skills during history taking and physical examination, attitude when interacting with the patient and skills during the whole period of doing clinical examination.

S/N	Component	Located marks	Attained score	Total section score	Comments
1	**ATTITUDE DURING EXAMINATION**			**6**	
	Appearance, dressing code and preparation (equip)	2			
	Introduced him/herself to the patient and informed the patient about the aim the procedure	1			
	Ensured patient/client privacy and confidentiality	1			
	Obtains informed consent	1			
	Appropriate patient positioning; ensure safety and comfort	1			
2	**HISTORY**			**16**	
	Asked about the chief complaints and duration in chronological order	2			
	Amplification of the History of presenting illness well detailed and explore appropriate information about the illness,	5			
	Review of other systems	2			
	Past medical history				
	• Prenatal	1			
	• Natal	1			
	• Post natal	1			
	Dietary history	1			
	Immunization history	1			
	Developmental milestones	1			
	Family and social history	1			
3	**PHYSICAL EXAMINATION**			**25**	
	Perform general examination appropriately	4			
	Perform systemic examination appropriately				
	Points to be added in the affected system	3			
	Cardiovascular system	4			
	Respiratory system	4			
	Per abdomen	4			
	Nervous system	4			

	Interpret signs	2			
4	**Summary and Presentation skills**			**8**	
	Summary	3			
	All information obtained from the history well presented	3			
	All information obtained from physical examination well presented	2			
	Deduction of marks if reported findings not asked or done during history taking or physical examination	-2.5			
5	**DIAGNOSIS**			**10**	
	Provisional diagnosis (Narrates features in support and against)	5			
	Provides and defends differential diagnoses	5			
6	**INVESTIGATIONS**			**6**	
	Provides the appropriate list and rationale of appropriate investigations	6			
7	**PATIENT CARE PLAN**			**14**	
	Discusses the most suitable treatment option according to the diagnosis/es				
	Correct medicine(s)	5			
	Correct dose	3			
	Correct schedule	3			
	Provided the follow-up plan and prognosis	3			
8	**PREVENTION**			**5**	
	Outline various modes of prevention	5			
	DISCUSSION IN RELATION TO PATIENT	**10**		**10**	
	Deduction for dangerous mistake **(Maximum deduction should be 10 marks)**				
	TOTAL SCORE			**100**	

NB: If the student reports something which he/she did not perform, half of the allocated marks for that specified area should be deducted. Likewise, if the student failed to perform a certain procedure during the course of history taking and physical examination in Serial Number 1, 2 and 3, the student will score 0 mark in that specified area.

Examiner's general comments --

Examiner's Name -- Signature -------------------------------- Date ---------------

THE UNITED REPUBLIC OF TANZANIA

MINISTRY OF HEALTH AND SOCIAL WELFARE

TECHNICIAN CERTIFICATE IN CLINICAL MEDICINE

RATING SCALE FOR CLINICAL EXAMINATION IN SURGERY

NAME OF SCHOOL --

Candidate's examination Number: --

The examiner shall observe closely when the student is demonstrating skills during history taking and physical examination, attitude when interacting with the patient and skills during the whole period of doing clinical examination.

S/N	**Component**	**Located marks**	**Attained score**	**Total section score**	**Comments**
1	ATTITUDE DURING EXAMINATION			6	
	Appearance, dressing code and preparation (equip)	2			
	Introduced him/herself to the patient and informed the patient about the aim the procedure	1			
	Ensured patient/client privacy and confidentiality	1			
	Obtains informed consent	1			
	Appropriate patient positioning; ensure safety and comfort	1			
2	HISTORY			15	
	Asked about the chief complaints and duration in chronological order	2			
	Amplification of the History of presenting illness well detailed and explore appropriate information about the illness,	8			
	Review of other systems	1			
	Past Medical/Surgical history	3			
	Family and social history	1			
3	PHYSICAL EXAMINATION			25	
	Perform general examination appropriately	4			
	Perform systemic examination appropriately				
	Local (if no local area affected, the points go to affected system)	4			
	Cardiovascular system	4			
	Respiratory system	4			
	Per abdomen (plus Digital rectal examination)	4			
	Nervous system	3			
	Interpret signs	2			
4	**Summary and Presentation skills**			8	
	Summary	3			

	All information obtained from the history well presented	3			
	All information obtained from physical examination well presented	2			
	Deduction of marks if reported findings not asked or done during history taking or physical examination	- 2.5			
5	DIAGNOSIS			**10**	
	Provisional diagnosis (Narrates features in support and against)	5			
	Provides and defends differential diagnoses	5			
6	INVESTIGATIONS			**6**	
	Provides the appropriate list and rationale of appropriate investigations	6			
7	PATIENT CARE PLAN			**15**	
	Discusses the most suitable treatment option according to the diagnosis/es				
	Correct medicine(s)	6			
	Correct dose	3			
	Correct schedule	3			
	Provided the follow-up plan and prognosis	3			
8	PREVENTION			**5**	
	Outline various modes of prevention	5			
	DISCUSSION IN RELATION TO PATIENT	**10**		**10**	
	Deduction for dangerous mistake **(Maximum deduction should be 10 marks)**				
	TOTAL SCORE			**100**	

NB: If the student reports something which he/she did not perform, half of the allocated marks for that specified area should be deducted. Likewise, if the student failed to perform a certain procedure during the course of history taking and physical examination in Serial Number 1, 2 and 3, the student will score 0 mark in that specified area.

Examiner's general comments --

Examiner's Name -- Signature -------------------------------- Date ------------