2019

EXPLORING THE VALIDITY OF MULTIMEDIA WRITTEN ASSESSMENT IN SAUDI ARABIA

Kattan, Thuraya Essam A

Copyright Statement

UNIVERSITY OF
PLYMOUTH
Doctoral College

EXPLORING THE VALIDITY OF MULTIMEDIA WRITTEN ASSESSMENT IN SAUDI ARABIA

by

THURAYA ESSAM KATTAN

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

Peninsula Medical School

October 2019

Acknowledgements

Foremost, I would like to thank my God ‘Allah’ for blessing me throughout this journey and for giving me patience. Undertaking this PhD has been a growing life-changing experience for me and my family and would have not been possible without the support and guidance from Allah, my family, colleagues, and friends. I would like to offer special thanks to the late Professor. James Ware my supervisor who although no longer with us has made this journey possible by believing in me. I would also like to express my sincere gratitude to my supervisors Professor Julian Archer, Dr. Lee Coombes, Professor Mohi Eldien Magzoub and Director of Studies Dr. Thomas Gale, for all the support and encouragement they have given me. Without their guidance and constant feedback and effort, this PhD would not have been achievable.

My sincere thanks also go to the Saudi Commission for Health Specialties for sponsoring my project and all the staff who during this long journey has made this research possible. I am also very grateful to Prof. Sulaiman Al-Emran, Dr. Mohammad Al-Sultan, Prof. Abdulaziz Al-Saigh, Dr. Yasser Alaska, Prof. Ayman Abdo, Prof. Larry Mellick, Prof. Larry Stack, Dr. Peter Devitt, Prof. Waleed Murani, Dr. Bechara Ghorayeb, Dr Mohamed Haddoud, Dr. Ahmed Mohammed, Dr. Amir Omair, Dr. Dimitry Abbakumov, Dr. Khaleel Al-Harbi, Prof. Mustafa Podrick, Dr. Majed Al-Salamah, Dr. Raed Hejazi, Dr. Nawfal Aljerian, Dr. Badr Alotaibi, Dr. Khalid Al-Rajhi, Stephen Fouchard, Nureldin Ibrahim, Saeed Alenezi, Dr. Maha Iqbal, Dr. Amani Al-Muallem, Dr. Hadeel Al-Kofide, Dr. Rabab Al-Kutbi, Luma Al-Abdulwahid, and the Saudi Board of Emergency Medicine Exam Committee (2012-2016), who were continuously helpful throughout my research and provided their assistance, resources and experience. I also want to thank Bernice Wilmshurst and Francesca Niedzielski for their continuous support throughout my studies. Special thanks for Nicola James

and all the staff at Bambino Child Care Centre, and Alison James for their care, support and hospitality in the UK and for making me and my family feel at home.

I am especially grateful to my Mother, Father, husband and son for supporting me emotionally, spiritually and educationally throughout my life and having love and patience to see me through this journey. I am also thankful and grateful to my brothers, sisters, and my mother-in-law and family for believing in me, supporting me and providing me endless encouragement. Finally, I would like to extend my thanks to all my colleagues and friends for all the ups and downs and fun we had and for their care, and to anyone who helped me directly or indirectly throughout my study.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee. Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment. This study was financed with the aid of a studentship from the sponsor Saudi Commission for Health Specialties.

A programme of advanced study was undertaken, which included:

EBM statistics, Excel 2012, Research owning & Using, Writing your literature review, Presenting at conferences, Overview to searching & accessing information resources, Introduction to qualitative research methods, SPSS, Referencing & paraphrasing, Introduction to Endnote, Presenting to an audience, Research methodologies, Effective reading, Originality & plagiarism, How to construct good quality MCQs, Innovative concepts in health professional education, Psychometric evaluation, Assessments modules of Master's program, Introduction to inferential statistics, Educational measurement, IRT, conducting a literature review, Advanced assessment course, Validity assessment, Assessment of clinical reasoning, OTTAWA – ICME.

The following external institutions were visited for consultation purposes:

National Centre for Assessment (Qiyas), Riyadh; National Guard Health Affairs (Medical Education Dpt.), Riyadh;
Kuwait University; Kuwait.

Presentations at conferences:

PU PSMD Research Event, Plymouth University 7th March 2013 (Oral Presentation)
PU PSMD Research Event, Plymouth University 22th October 2014 (Poster presentation)
PU PSMD Research Event, Plymouth University 14th April 2015 (Poster presentation)
PhD showcase: Shaping Cultures, Plymouth University 25th Jan 2018 (Oral Presentation)

Word count of main body of thesis: 95232

Signed: Thuraya Kattan

Date Oct 15, 2019

Abstract

Introduction:

Multiple choice questions are widely used in high-stakes written examinations and are continuously being challenged for testing recall facts rather than higher cognition. Innovative MCQ formats (Multimedia-enhanced) can test such skills. To evaluate this format a recent validity framework “The Cambridge Framework” was used to assess and explore this intervention.

Aim:

Explore the validity of multimedia MCQs for testing higher cognition in Emergency Medicine (EM) in Saudi Arabia using the Cambridge framework, and evaluate the use of this framework.

Methods:

A total of 164 EM residents (seniors and juniors) from three regions of Saudi Arabia took a total of 80 multimedia-text matched items in an end of year exam. A mixed-methods approach triangulating quantitative (pilot test, parallel forms of multimedia and text items, item psychometrics and characteristics, questionnaires) and qualitative methods (semi-systematic literature review, focus-group discussions, Cambridge validity framework implementation and research legitimation), were applied using systematic guidelines for each method to explore multimedia items.

Results:

Discrimination was significantly higher for multimedia than text items (DI= 0.19 $\pm$18, 0.14 $\pm$17, p= .03), and took significantly longer to answer (p=.01). Both formats had a moderate difficulty level (Diff = 0.75,0.74). Multimedia-items had a higher reliability and G-coefficient than text items. Focus group results revealed seven main themes of multimedia that effect item characteristics. Review of the Cambridge Framework

demonstrated areas of gaps in sources of validity evidence and external-related factors not covered in other frameworks that have implications on validity.

Conclusion:

Multimedia questions were more discriminating and took longer time to answer than the text questions. They test higher cognition and have certain characteristics that effect difficulty level and how they are perceived by the examinees altering their thinking process towards the answer. Gap areas identified in the Cambridge framework and external factors that were contextual, may give room to explore the issue of international validity in assessment.

Table of contents

List of tables

List of Figures

List of Abbreviations

AERA	American Educational Research Association
AMEE	Association of Medical Education in Europe
APA	American Psychological Association
AV	Audio-visual
BP	Blueprint
CAT	Computer-adaptive testing
CBT	Computer-based testing
CCS	Computer-based case simulation
CIV	Construct irrelevant variance
CK	Clinical Knowledge
CLT	Cognitive load theory
CST	Clinical simulation test
CTT	Classical test theory
DI	Discriminating index
DIF	Differential Item Functioning
DIFF	Difficulty level
DV	Dependent variable
ED	Emergency Department
EM	Emergency Medicine
EMQ	Extended matching questions
FG	Focus group
GRE	Graduate Record Examinations
IA	Item analysis
IRT	Item response theory

ITC	International Test Commission
IUA	Interpretive use argument
IV	Independent variable
IWF	Item writing flaws
LTM	Long-term memory
MM	Multimedia
MCQ	Multiple-choice questions
MIPP	Multimedia integrated Pilot Project
NBME	National Board of Medical Examiners
NCLEX	National licensure examinations for nurses
NCME	National Council on Measurement in Education
OSCE	Objective Structured Clinical Examinations
PMIT	Perceptually motivated item types
PMP	Patient-management problems
PNP	Paper-and-pencil
QUAN	Quantitative
QUAL	Qualitative
RCPSC	Royal College of Physicians and Surgeons of Canada
RPB	Point biserial discrimination
SBA	Single best answer
SCFHS	Saudi Commission for Health Specialties
SEM	Standard error of measurement
SS	Standard Setting
STP	Secure Test Portal
TCA	Test centre administrator
TEI	Technology-enhanced items

TSD	Test specification document
TXT	Text
USMLE	United States Medical Licensing Examination
WM	Working Memory

Appendices Headlines

Appendix 1 Variables related to the Research Question

Appendix 2. Summary of the 11 Studies that used MM-TXT matched items

Appendix 3: Examples of proposed Frameworks and Dimensionalities for Multimedia Classification

Appendix 4: Combined view of Bennett et al. (1999); and Lui et al. (2001) for the development of multimedia in assessment

Appendix 5: Types of validities

Appendix 6 Outline of the overarching standards from AREA, APA and NCME (2014)

Appendix 7: Examples of Validity Frameworks

Appendix 8: The 12 components (steps) for an Effective Test Development Process

Appendix 9: Strengths and Weakness of Mixed Method Research

Appendix 10: Principles of Questionnaire Construction

Appendix 11: Study Questionnaire

Appendix12: Information sheet and consent form

Appendix 13: Points checked by Specialist and Content expert reviewers for items

Appendix 14: CBT exam specification

Appendix 15: Strength and weaknesses of MCQs

Appendix 16: List of possible MM-TXT topics

Appendix 17: Test Specification Document (TSD)

Appendix 18: Cambridge Framework

Appendix 19: Checklist for evaluating a Mixed-Method Research Study

Appendix 20: Validity checks for quantitative, qualitative and mixed methods research

Appendix 21: Example of a combined result of QUAN and QUAL analysis

Chapter 1: Introduction

1.1 General introduction

This thesis focuses on one of the widely used methods in assessment and its application in postgraduate medical education: the use of multiple-choice questions (MCQ) in assessing residents' higher cognitive skills, one of the many competencies that are required by medical and health graduates in Saudi Arabia.

This chapter focuses on the background of the study regarding the multiple-choice assessment of postgraduate medical specialities in Saudi Arabia and is based on the researcher's own experience as assessment head in the department of medical education and postgraduate studies in the Saudi Commission for Health Specialties (SCFHS) in Saudi Arabia. The chapter will start by stating the problem and the research questions, then go on to introduce the idea of including multimedia (MM) into written examinations in order to assess the graduates' higher-order thinking skills. It will conclude by outlining the research methods, as well as an explanation of the thesis structure.

1.2 Statement of the problem and research questions

This research sought to firstly establish a better undertaking of the literature to fully inform the development of new multimedia questions for high-stakes examinations in Saudi Arabia. An Initial search using related search terms of multimedia, MCQs, and written examination revealed 21 articles of which only four were relevant to the topic. These initial results seemed to imply that there was a gap in the literature covering this area and that further exploration was needed to be able to understand the characteristics of multimedia, as well as its effect on students' performance on test results. The work, within this thesis, explores with emergency medicine residents in the national programs of Saudi Arabia, the use of multimedia in an end of year examinations against a new validity framework as a precursor to introducing the

concept into national licensing examinations. This would hopefully aid in illuminating certain areas in test development by answering the following question: Can the use of multimedia in MCQ examinations test higher cognitive skills (the construct) that is more than rote memorization in postgraduate residents? And is it more appropriate than traditional MCQs? To help in answering this, the following research questions were explored:

1) Does multimedia-enhanced MCQ's test higher cognitive levels more than text questions in high-stakes examinations? In order to answer this question, the following sub-questions also need to be answered:

   Questions related to the items

   - What is meant by higher cognitive skills?
   - Can MCQ measure higher cognitive skills?
   - What are the psychometric properties of multimedia questions?
   - Do multimedia-enhanced MCQs produce higher psychometric properties than traditional text MCQs?
   - What are the characteristics of multimedia in assessment, what factors affect it, and when and how should it be used?

   Questions related to the participants

   - What are the participants' perceptions of using a computer-based setting for their examinations?
   - What are the participants' perceptions of the use of multimedia items in their examinations?

The research question is further fragmented in the form of variables (dependent and independent) and is explained with examples in Appendix 1.

How can it be made sure that the interpretations of the results are valid? This can be answered through the application of a validity framework, and in order to answer this

question, the following one needs to be addressed:

2) Can the new Validity Framework (Cambridge Assessment Group) be used for evidencing assessment validity in large-scale high-stakes examinations in a different setting (multimedia MCQ examination" in Saudi Arabia, as the central part of the original research)? To answer this the following questions regarding test development, need to be asked:
    - What does this validity framework cover, and how does it compare to other frameworks?
    - How easy/difficult was it to understand and apply the framework?
    - What are the challenges and gaps when using this new framework?

It is hypothesized that because multimedia is more complex than text description then multimedia items would be more difficult to answer than text items. It is assumed that these questions would require more cognitive processing to answer and, consequently, would produce test results that are more discriminating between low and high-ability students. Therefore, the following hypotheses were claimed:

HYPOTHESIS 1: Residents who took the multimedia MCQs would have a lower mean test score (i.e., items are more difficult) than those who took the description (text) format.

HYPOTHESIS 2: Multimedia MCQs would have a higher discrimination index than text MCQs.

1.3 Background and study rationale

In the past two decades, the Kingdom of Saudi Arabia has experienced a sudden expansion in the healthcare system, with a sudden increase in numbers of medical schools in a short period of time from seven to 33. But even with the increased number of medical graduates yearly, Saudi Arabia still falls short in meeting the

medical manpower needed for a growing population of around 32 million people. One way to meet these needs for the public healthcare system is to certify and/or license physicians locally, as well as outside the country until sufficient local physicians are attained.

In the field of healthcare, assessment should be able to determine if a candidate would be able to become a doctor who has obtained the knowledge and skills to take care and manage a patient (1). Health practitioners are faced on a daily basis with cases of complexity and a great deal of information and they are required to arrive at the right decision for patient care. These complex cognitive abilities need to be tested throughout their training to ensure that they have the capability to utilise all available information to make decisions when practicing in the clinical environment (2). Summative assessment should determine which resident is fit to proceed to the next level and to be able to practice autonomously (1). The quality of high-stakes examinations in the Kingdom, the Middle East, as well as other international licensing examinations around the world, is mainly focused on testing knowledge using text and static-image based MCQs when applicable (2-4). In Saudi Arabia, a big part of assessment relies heavily on written examinations, particularly multiple-choice examinations, and it would take some time until the introduction of a formal Objective Structured Clinical Examinations (OSCE) system or other forms of clinical examinations become efficiently established and standardized for high-stakes medical specialty examinations.

The SCFHS is the regulatory body responsible for training, certifying, and licensing physicians to practice in the Kingdom of Saudi Arabia. In response to public demands to provide competent, safe, and fit-to-practice doctors, the Commission has worked to raise the standards of education, training, and assessment to those of international levels. With over 70 medical licensing examinations, and 62 board

certifying examinations that are conducted yearly, the Commission has decided to review and upgrade these examinations through a newly established Medical Education Department and with the collaborative support of international sister-like organisations.

In the last decade, the Saudi Commission has been reforming assessment methods to include new innovative methods that meet international medical education standards, and one of their physical transformations was the construction of in-house computer labs for testing in all of their branches in the Kingdom. In the meantime, there is a need to establish a well-constructed written examination with clear standards that test not only knowledge but potentially other higher-order competencies (5). One possible way of designing such an exam, which would focus on problem-solving and application of knowledge rather than pure memorization of facts, would be the use of innovative items, such as multimedia-enhanced MCQs (6, 7).

Innovations for assessing learning in medical education over the last half a century have challenged traditional approaches, such as the MCQ examination. Newer formats, such as OSCEs (8), Workplace Based Assessment (WBA) (9), and, most recently, Situational Judgement Testing (SJT) (10), all claim to be testing additional constructs to core medical knowledge. However, there is convincing evidence that knowledge, a common educational objective (11) is the single best determinant of expertise in practice (12-14) and that cognition is best tested using written examinations (15).

The move towards other testing formats has been driven by the desire to test higher-order cognitive function, such as decision making and clinical reasoning (16), but as importantly, there has also been a drive to support perceived examination fairness and relevance. MCQs have been criticized for not being fair because they

promote short-term superficial, rather than deep learning (17) and lack authenticity for clinical practice (18). Fairness is important as it equates to defensibility, which has become increasingly important in modern assessment (13). In response to these claims, educational testers have explored ways of reducing 'cram' learning using more regular testing formats and exploring a range of ways of using new technology-enhanced learning techniques to support professional authenticity (5, 19, 20).

MCQs have been used in assessment for over 45 years in medical, dental, nursing, and other healthcare disciplines in the US, UK, Europe, and Australia (2, 21). However, educators use MCQs to test factual knowledge rather than to test a deeper understanding (21). They are widely accepted as a user-friendly strategy for assessing knowledge throughout different educational systems and disciplines and are seen as being a reliable, valid, and efficient way to test learning outcomes (22). It is traditionally known that MCQs predominate certificate and licensure examinations worldwide (22, 23), as they are considered the most commonly used assessment method of diagnostic reasoning. The bulk of USMLE step 1, 2, and 3, Part I of the Medical Council of Canada Qualifying examinations (24) and medical licensing exams in Germany (25) are MCQs. Written tests, including MCQs, are like any other educational methods and are not without fault. The advantages of written tests, including MCQs, are that they are practical to develop, feasible to deliver, and achieve high levels of reliability. However, much of their reliability is assured by being able to sample large amounts of content in relatively short periods of time (13). Despite concerns that MCQs test recall of isolated facts, there is evidence that supports the notion that MCQs will discriminate accurately between candidates and can be used to measure higher-order cognitive skills (26) (27 p.38), especially when written clearly and constructed well (13) and when using clinical

vignettes (2). MCQs with clinical vignettes are widely used in medical education (28) and are likely to be used as a summative method (4). They are used in high-stakes nursing assessments as well as other health science disciplines (2); however, students' clinical reasoning strategies haven't been well described in these questions (28). With the innovation of technology and the use of computer-based testing (CBT), it has become possible to introduce the use of multimedia and simulation into written examinations, raising the hierarchy of clinical testing still further (29) making the questions more authentic (30), and setting a realistic perspective of what takes place in clinical practice (23). Although the use of images has become important, its interpretation is still not well documented (31). Complex items, such as well-written clinical vignettes, rely on the use of knowledge (32 p.126), and with the use of multimedia (complex material), examinees are required to use higher mental processes (32 p.126). In addition, one of the primary uses of multimedia materials is to complete the presentation of the presented problem that is being solved (33 p.94). A good multiple-choice item requires more than recall to be answered (32 p.127) as examinees are asked to interpret an image or video that requires their application of knowledge and hence, MCQs can measure any aspect of cognition (32 p.132).

When talking about decision making and problem-solving, one tends to forget that it is not a simple process. It is an abstract of the mind that cannot be seen and is not a neat straight-forward logical pattern. It is more complex because, in the real setting, the physician is not only dealing with the thinking process. He is also dealing with involving others in the decision-making process, having inadequate information, feelings of uncertainty and being restricted in time to make a decision; and with all this, one could not know which determined his/her behaviour or even ask them to explain why they acted in a particular way (34). In such settings, the use of high and

low fidelity multimedia and simulation can be used to construct such scenarios, where ‘real-life’ deliberate thinking of students, residents and physicians can be assessed (34). Assessment of clinical reasoning can be achieved through the use of testing modalities, such as patient management problems, script concordance tests, key feature examinations, and think-aloud protocols (28). However, they are not frequently used because of their required resources and inadequacy in testing (28). Nevertheless, MCQs with clinical vignettes, are efficient enough to be administered in examinations and yield excellent psychometric properties (28).

In order to be able to evaluate such a major change (introducing multimedia) in a high-stakes setting of national examinations, it is important to have a robust approach to assess the process of test development taken. There are well-established validity frameworks with which to explore assessment interventions. This research drew on such frameworks that are discussed in more detail in the coming chapters. One of these frameworks was very new at the start of this research and claimed to ‘operationalize’ validity evidence for large-scale international examinations (35). This framework was implemented to validate the use of multimedia multiple-choice questions (MM-MCQs) in the examination, and the framework was evaluated for its applicability in a different setting (Saudi Arabia) as the central part of this research. The research takes a ‘consequentialist’ stance drawing on the work of Kane (36) and building on an important newly proposed framework for evidencing assessment validity in large-scale, high-stakes examinations “The Cambridge Validity Framework” (35). Kane argues that we often only focus on scores and their interpretation in assessment. He postulates that we need to think as much about the consequences of these interpretations as we do the interpretations themselves (36). The mixed-methods approach of the Cambridge Validity Framework appears to be an appropriate modern assessment

evaluative approach that draws on the strengths of both qualitative and quantitative data collection. This study will explore a fuller understanding of the consequences of a new assessment intervention "MM-MCQs" within the validity framework from the Cambridge Assessment, which was newly published at the start of this PhD, and will establish if this framework is fit for purpose for the implementation of a high-stakes end-of-year examination in the Middle East. Figure 1.1 depicts the main components of the study.

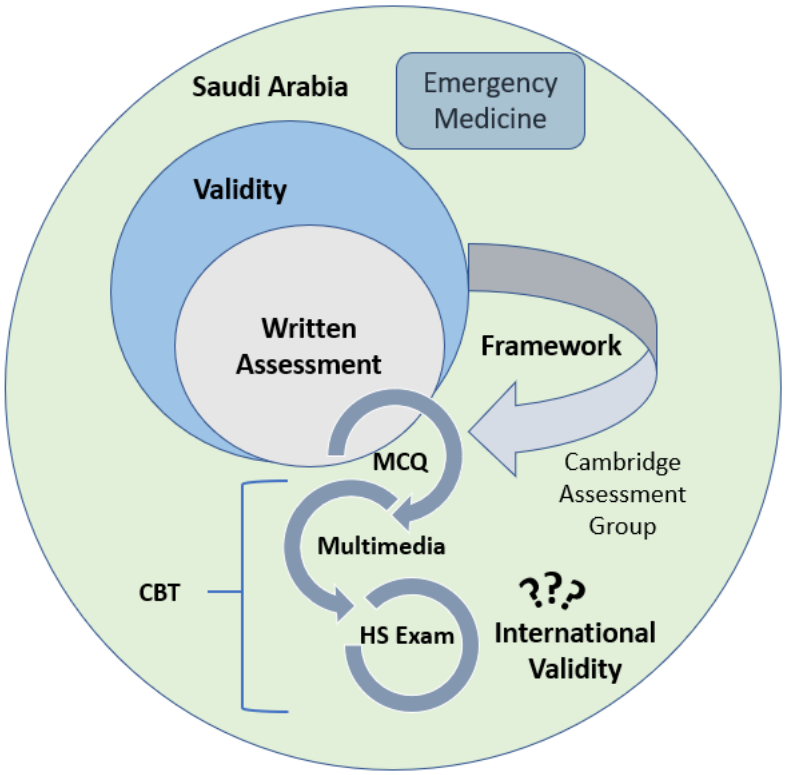


Figure 1.1: Main study components

As demonstrated in the image, the largest circle is the setting in which the study takes place "Saudi Arabia", in the speciality of emergency medicine. This would be conducted in their end-of-year written examination (high-stake), which is in the MCQ format. The MCQ examination would include multimedia items that would be compared to their text-matched items and, therefore, would be conducted through

computer-based testing (CBT). The whole test development of the examination fits into assessment validity because score interpretation from these items needs to be validated in this case, using the Cambridge assessment group validity framework.

The following table (Table 1.1) outlines a series of quantitative and qualitative methods used to address the research questions, how they were carried out (methodology), and the analyses that were undertaken.

Table 1.1: Outline of research questions, methods, methodology, and data analysis used

Main Research Questions

1. Do multimedia-enhanced MCQ's test higher cognitive levels more than text questions in high-stakes examinations?
2. Can the new Validity Framework (Cambridge Assessment Group) be used for evidencing assessment validity in large-scale high-stakes examinations in a different setting (multimedia MCQ examination" in SA, as the central part of the research)?

Research Methods	Related Research Sub-Questions	Methodology (approach used)	Data analysis and Justification
1.Literature review (Qualitative)	What are the psychometric properties of multimedia questions? Do multimedia-enhanced MCQs produce higher psychometric properties than traditional text MCQs? What are the characteristics of multimedia in assessment, what factors affect it, and when and how should it be used?	Semi-systematic literature review of the health professionals and general educational literature, using appropriate search engines and databases. The PRISMA checklist for systematic reviews was used as a guide for the research questions. Search term strategies and PICO were used to analyse the question.	A narrative review of the articles from the literature review summarising the main findings in the literature and interpreting evidence related to the research question. Reporting previous studies on MM used within written assessment and in MCQs, the evidence and methods especially, but not exclusively, in high stake examinations across the professions. This will shed light on available guidelines to help set up the MM exam items, as well as highlight areas of gap in the literature that need addressing using other methods, and that can be applied to this research.
2. Pilot Study (Quantitative)	What are the psychometric properties of multimedia questions? Do multimedia enhanced MCQs produce higher psychometric properties than traditional text MCQs?	The pilot study was conducted by developing multimedia-text matched questions and administering them in a real-EM examination through computer-based testing (CBT). Guidelines for test development, CBT, validity frameworks, as well as the *Standards,* were used.	Reporting the details of the experience, process, and outcomes, initial analysis of the item characteristics through item psychometric analysis of the new multimedia questions. This provides an opportunity to identify the feasibility of the project, possible barriers and to test any areas requiring revision in the proposed protocol involving the various stakeholders. The pilot study draws from results of the literature and focuses, on methods 3, 4, and parts of 6 that were explored.

3. Tests: Form Development (Quantitative)	What is meant by higher cognitive skills? And what measurable features does it contain? Can MCQ measure higher cognitive skills?	Using evidence found in the literature, two forms of examination were developed: (a) MCQs with multimedia in them and (b) a parallel text-matched question with no multimedia in it. This will be administered as unmarked beta questions in the Emergency Medicine (EM) end-of-year examination twice, once in the pilot phase and next in the study. CBT, validity frameworks, as well as the *Standards* were used.	Data on sample demographics and characteristics, item indices and characteristics (P-value, item discrimination, difficulty, item-total biserial correlation, and response time) using the classical test theory will be collected. Further data exploration on reliability using Cronbach alpha and generalizability theory will be conducted. Data analysis using SPSS, Minitab, Excel was carried out to help identify multimedia item characteristics, as well as a guide to item selection for group discussion in the focus group These will differentiate between MM-TXT items, their characteristics, and effect on the construct.
4. Questionnaire (Quantitative)	What are the participants' perceptions of using a computer-based setting for their examinations? What are the participants' perceptions of the use of multimedia items in their examinations?	Computer-based confidential and anonymous Questionnaire targeting Saudi EM residents. This will be circulated after the examination to gain an understanding of their experience, concerns and acceptability of the use of multimedia questions, as well as to improve their use.	Using SPSS and Excel for calculating and reporting correct response rate and group comparisons and reporting on descriptive statistics. This will complement finding and augment results from test results, and aid in the focus group discussions and results.
5. Focus Group (Qualitative)	What are the characteristics of multimedia in assessment, what factors affect it and when and how should it be used? What are the participants' perceptions of using a computer-based setting for their	Conducting focus groups in three regions in Saudi Arabia with the EM residents to explore their collective view of the new items. Using data from the questionnaire, as well as the MM-TXT matched item analysis to guide the discussion. A systematic focus	Transcribing the discussions using NVivo system. Using a six-phase thematic analysis guide for the process of transcribing, coding, and outlining emerging themes. This is to evaluate the acceptability of the new items through the qualitative analysis, highlight multimedia features and challenges, as well as item behaviour in relation to analysis. This will

	examinations? What are the participants' perceptions of the use of multimedia items in their examinations?	group conduction through guidelines were used covering participant recruitment, selection, material and session preparation, moderator role, group interactions, and data analysis to carry out the focus group discussion.	aid in gathering rich qualitative data in a short time and complement, elaborate, and strengthen results from the other methods (item analysis results from tests and questionnaire results).
6. Validity Framework (Qualitative)	How easy/difficult was it to understand and apply the framework? What are the challenges, gaps when using this new framework? Was the framework applicable in this new setting? Can it be applicable elsewhere in other settings?	Using the Cambridge framework, a modern assessment evaluative approach that draws on the strengths of both qualitative and quantitative data collection to evaluate the results of the MM-TXT examinations (pilot and test).	Analysis in this section is through a narrative report going step by step through the framework to evaluate the applicability of the Cambridge framework in establishing the validity of the MM-TXT examinations in Saudi Arabia. Reporting on all its steps, and methods and providing sources of validity evidence by identifying the strengths and weaknesses to validate the interpretations of the IA results. In addition, highlighting the challenges and gaps in the framework. This framework will help validate results from the test, identify areas of improvement, and help explore the concept of international validity through its application in a different setting (Saudi Arabia).
7. Legitimation (Qualitative)	How can we make sure that our interpretations of the results are valid?	Legitimation (i.e., validity in mixed-methods research) covers different types of validities in quantitative, qualitative, and mixed-methods research that one should review in order to check for the quality of the research that was conducted. Onwuegbuzie and Collins' sampling framework was used, as well as a checklist to evaluate the quality of a mixed research study.	Evaluating and reporting on the items on the checklist, as well as reporting and providing examples regarding the multiple validities proposed by Onwuegbuzie and Collins to demonstrate the validity of the research process and its results. Reflecting on the whole research study and the different types of validities applied to mixed research, explaining strengths, weaknesses, and limitations of the research.

1.4 Significance of the research question

This work is viewed as important to Saudi Arabia as this research would inform and develop the assessment systems in postgraduate examinations and as a second step in the Licensing Examinations. The researcher's position within the organisation would ensure that this research would lead to real policy development and change.

The work is central in informing the wider literature by exploring the applicability of a new validity framework. The research would be theoretically sound, potentially building new theory through action. The importance of raising the quality of high-stakes examinations is of global interest. Results of this study through dissemination would find ways to implement the newly proposed method and thus enhance high-stakes examinations globally in relation to the use of multimedia multiple-choice formats in high-stake examinations. The introduction of multimedia to written and clinical examination would help raise the quality of practising doctors by raising the assessment standards of clinical competence in written examinations. Developing an understanding of validity is of importance to the wider educational literature. It is hoped that results from analysing the application of the Cambridge Validity Framework in a Saudi context will yield evidence in exploring the possible existence of a new concept in validity "international validity". This would be by evaluating if the framework was applicable in a new setting as it claims to be and whether it would be applicable elsewhere. In addition, the research aims to report on the main psychometric findings from the study and on the validity evidence for the use of multimedia items in written examinations, as well as highlight the characteristics of multimedia items needed in an examination.

1.5 Thesis outline and summary of chapters

This research aims to answer the research questions through the proposed mixed-methods outlined in Table 1.1 and will be discussed in seven chapters.

1.5.1 Chapter 1: Introduction

This first chapter as discussed here provides a brief overview regarding the research background, as well as the research problem. It goes through outlining the research questions and related sub-questions, outlining the proposed methods and methodologies used, as well as the structure of the research.

1.5.2. Chapter 2: Literature review

This second chapter explains the literature review which is the first method listed in Table 1.1. This chapter explains the justification for conducting a literature review on multimedia in multiple-choice examinations. The chapter then explains the approach taken to conduct a semi-systematic literature review, explaining the search strategy, search terms, criteria and databases used. Further on, results and discussion of the outcome of the research are explained.

1.5.3 Chapter 3: Multimedia literature review

Based on the results of the literature review in Chapter 2, this chapter covers of what was found on multimedia in MCQs examinations. It starts by introducing computer-based testing (CBT) and its use in assessment in medical education followed by the introduction of multimedia and authentic assessment. The chapter goes on to discuss the historical background of multimedia in testing and in MCQ written examination, multimedia learning, and the principle of cognitive load theory.

1.5.4 Chapter 4: Validity and validity framework

This chapter covers the complicated concept of the construct, validity, and validity frameworks in the assessment and test development process providing a brief history on validity, as well as examples of some of the well-known validity frameworks that are cited in the literature and that are used in the test development process. Further details of each framework and its concepts are also outlined.

1.5.5 Chapter 5: Methodology

This chapter covers the research design and methodology carried out to conduct the study. In this chapter, the qualitative and quantitative research methods are explained and include the conduction of the pilot study, the development of the multimedia text-matched items and questionnaire, focus group conduction, the use of the Cambridge Assessment framework and its implementation during this research, as well as research legitimation.

1.5.6 Chapter 6: Results

This chapter presents the analyses and results' findings from the various methods used in this study and covers two sections: (a) quantitative analysis and (b) qualitative analysis. The quantitative analysis covers results outlined from the exam item analysis, as well as the questionnaire analysis while the qualitative analysis covers the transcribed data from the focus group, as well as qualitative analysis of a sample of MM-TXT matched items selected and the validity framework and legitimation.

1.5.7 Chapter 7: Discussion

This chapter draws on the main results of the quantitative and qualitative data drawn out from the study and converges them into an explanation of why the items behaved as they did. It also explains the outcome of the framework and discusses the concept of international validity. It ends with a reflective section from the researcher's point of view, as well as recommendations when using multimedia materials in examinations.

1.6 Conclusion

Innovations for assessing learning in high-stakes examinations have challenged traditional approaches, such as the multiple-choice question to establish new testing formats that test higher-order competencies (e.g., multimedia-enhanced MCQs). To evaluate this format in a high-stakes setting, a recent validity framework "The Cambridge Framework" that draws on the work of Kane was used to explore the validity of multimedia MCQs in the Saudi emergency medicine end-of-year examination.

Triangulation of quantitative and qualitative methods was carried out for this project and involved a semi-systematic literature review, pilot test, the use of parallel test forms of multimedia and text items, item psychometrics and characteristics analysis, questionnaire distribution, focus group discussions, validity framework application, and evaluation and research legitimation. These are all discussed in the following chapters.

Chapter 2: Literature Review

2.1 Introduction

This chapter addresses the first method used that was outlined in Table 1.1 to address the research project (i.e., literature review). The aim of this chapter is to explore the evidence for the use of multimedia in MCQ examinations in medical and healthcare settings. This chapter aims to address the first research question “Does multimedia-enhanced MCQ’s test higher cognitive levels more than text questions in high-stake examinations?” through trying to answer the following sub-questions:

- What are the psychometric properties of multimedia questions?
- Do multimedia enhanced MCQs produce higher psychometric properties than traditional text MCQs?
- What are the characteristics of multimedia in written assessment, what factors affect it, and when and how should it be used?

An explanation for the justification for using multimedia in this PhD research will first be explained followed by outlining the methodology used to carry out the literature review search and ending with the results, discussion, and limitation of this search.

2.2 Justification for using multimedia (MM) in written MCQ examinations

One of the main sources of validity evidence from assessment content is the extent to which its inclusion in assessment is relevant to the construct that is of interest (37). It is suggested that valid assessments that resemble real-life situations within the limitations of a standard testing condition would result in a valid meaningful learning experience (18). Computer animations and illustrated phenomena are increasingly being used in medical education to aid in understanding complex and abstract concepts (38). They have been used in clinical teaching in numerous medical

specialities to facilitate learning physical examinations, and various procedures and techniques (38). With the growth of computerized testing, students are expected to learn and process various materials in a variety of ways (printed and visualized) and educators are realizing the importance of evaluating students not only on the content of their materials but also on how they reason with it (39). Undergraduate medical programs entail students to be able to identify and interpret images (40). In many postgraduate medical disciplines, such as radiology, histology, pathology, cardiology, surgery, orthopaedics, emergency medicine (EM) and nursing, visual skills (such as interpreting radiographs, microscopic images, and reading electrocardiograms) and auditory skills (such as heart sounds) are considered essential skills for physicians to be competent (41-45). Residents are required to learn problem-solving skills and interpret findings in order to be able to arrive at a diagnosis, prognosis, or management plan. Clinical medicine is a discipline that would prosper well with the use of multimedia items (44). The use of simulation (high and low fidelity) in EM is considered important, as it delivers new features that cannot be captured with the conventional paper-and-pencil examination, such as creating the sense of urgency, forcing residents to think in a time-dependent manner, prioritizing care and making decisions given the limited information and time presented to them (46). This mirrors what actually happens in the emergency department (ED) setting. In addition, the use of simulation in EM has expanded since the late 1990s and has been a basis for many EM training programs (47, 48). As in the speciality of anatomy, many resources in emergency medicine have a visual component to it because the speciality is derived from almost all the other medical specialities, including anatomy (49, 50). However, a large component of the EM assessment, like other specialities, takes an approach away from visualization (49). Because visualization represents an important aspect of emergency education,

the effectiveness of multimedia (images and videos) used for assessment purposes requires further continuous study.

Moreover, in good educational practice, educators should construct assessment to match not only the educational objective (to ensure that content validity is present) (51) but to also align with the curriculum, its intended outcomes, teaching methods, and assessment tasks with each other (52). If tests were designed to measure key concepts and constructs that are taught in the medical curriculum but at the same time ignored these same concepts because they were harder and more laborious to measure, then educators would reduce their teaching in these areas (53), and it would be the test developers who would bear some of these responsibilities (53). There would also be a lack of validity if the teaching environment, the technology, and the use of multimedia were established well enough, but the assessment area still lagged behind and was based on the paper and pencil format (53). It is technically possible to include images into written paper-based examinations as it is including videos in computer-based examinations, but information regarding their effect on item statistical properties are limited (40, 54, 55). In addition, what is available in the literature from previous methods of including multimedia in the paper-based examination were of reducing images to simple diagrams, having low-quality images or providing exam booklets of illustrate colour plates (40, 55, 56).

In the past few years, the SCFHS have been upgrading their educational and assessment programs as part of the new vision of Saudi Arabia “Vision 2030”. The EM residency training program exam committee in SCFHS have updated their curriculum and required learning outcome for their residents and have made these explicit on their website (57) in order to be able to assess their competencies. Therefore, in order to warrant constructive alignment with authenticity, it was decided

that multimedia items would be included in their written examinations in order to test higher cognitive abilities (i.e., the desired construct under measurement in this study) (40, 58). Test results based on the interpretation of images and videos were used as a method for assessing residents' competency in their clinical decision-making skills.

However, much of what has been written in item-writing guidelines has been for MCQs, essays, and other written formats yet guidelines regarding multimedia and innovative items were rarely found (59). Moreover, at the start of this research, the two North American medical licensing organisations had not provided instructions or guidelines on the use of multimedia in their item writing instruction manual, although images do appear in their examinations (54, 60, 61). Though, recently in their updated book, the National Board of Medical Examiners dedicated a brief section with some recommendations for multimedia use (62), these updates did not have guidelines or instructions on what factors affected multimedia materials, and how the use of different multimedia materials affected test items and students' performance.

Furthermore, most studies have been conducted in countries, as Akdemir and Oguz (2008) put it, 'where the duration of technology integration process is short' (63). Reliable standards and guidelines are needed regarding integrating these technologies into assessment. Because multimedia plays an important role in our teaching and is among the most common feature in our training programs, exploring the influence of multimedia on item tests is a valuable starting point (54).

It is hoped that from the research results, accumulated instructions and guidelines from lessons learned throughout the literature would be applicable and would help test developers save time and effort when carrying out their multimedia examinations. The purpose of this research is not to replace previous ideas in the literature, but to gather,

organise, and encourage further conversations on this topic amongst educators and test developers.

2.3 Literature review: Semi-systematic review of multimedia in assessment

The following section describes the methodological process taken to search the literature and results obtained.

2.3.1 Introduction

A semi-systematic review of the literature that reports previous studies on the validity and consequences of including multimedia in MCQs was conducted. Initial scoping exercises were undertaken to assess the evidence within the health professionals and general educational literature to inform the review protocol and exact methods then employed. The purpose of the review was to have a better understanding of the use and validity of multimedia in written examinations especially but not exclusively in high-stake examinations across the professions. This aided in providing insights on the feasibility, logistics, and methods needed to be used for conducting the pilot project and, afterwards, the continuation of the study.

To ensure that the research methods used in this investigation were valid, reliable and appropriate for answering the study questions, a semi-systematic approach was carried out to review the literature. The PRISMA checklist (64) for systematic reviews was used as a guide for the research question. PRISMA stands for 'Preferred Reporting Items for Systematic Review and Meta-Analyses' and are an evidenced-based set of items that are required for reporting systematic reviews and meta-analysis (64). It is important to note that this search does not fulfil the criteria for a systematic review and was carried out by the researcher alone. The selection of

papers and data extraction was not reviewed by multiple authors as done in a systematic literature review. However, the usage of databases, search term strategies, and PICO were used as described in the following section.

2.3.2 Methodology

This section covers the approach taken for the semi-systematic literature review in the health professionals and general educational literature. It explains the strategies that were used to search terms to analyse the question, inclusion and exclusion criteria, as well as the appropriate search engines and databases that were used.

2.3.2.1 Search terms strategy

The first step taken to start this research was to dissect the research questions into simpler blocks that needed to be explored, this was followed by compiling a list of key terms and synonyms revealed from the literature that would aid finding the main related topics. The PICO strategy was used to analyse the question in order to be able to create this list (65) and is demonstrated in Table 2.1.

Table 2.1: PICO Strategy outlining the main research terms and their synonyms

PICO Acronym	Study Question
P (Patient, Problem, Population)	Medical education postgraduate and undergraduate students Including: - Medical students, - Dental students - Nursing - Health Allied students
I (Intervention)	MCQs with multimedia materials including: - Multimedia, multi-media, images, pictures, chart, graph, photo, photographs and illustrations - Video, animations, media, visual, virtual and simulation - Audio, sound, audio-visual and recording - Interactive, authentic, technology-enhanced and computer-based items - Multiple choice, MCQ
C (Control or Comparison)	MCQs with only text description including: - Text, written and description of an item
O (Outcome)	Changes in cognitive levels reflected by: - Item characteristics including test scores, Item difficulty, item discrimination and duration. - Students' perception towards the items including reaction, familiarity and authenticity felt towards the items

Regarding the selection of which publication would be included in this study, inclusion and exclusion criteria were determined (see Table 2.2) and all publications from 1960 to May 2018 (the time of the updated literature review) were included. This time frame was selected based on the following points:

1. The fact that computers were used for medical examinations since the 1960s to test knowledge and problem-solving skills by using MCQs and extended matching questions (EMQs) (1, 66).
2. Boulet and Swanson's article (2004) highlighted that patient-management problems (PMPs) have been used in examinations since the 1960s, and these

items require students to interact with the items in order to reach a response (67).

3. The use of film in large-scale testing by the US Army Air Force Aviation Psychology Program during World War II was documented to be in the 1960s (Siebert and Snow, 1965, cited (7)).

Most of the literature review was mainly concentrated on higher education in the medical and health allied field; however, it did consider those publications that were advanced in multimedia topics in other fields and educational levels. Inclusion and exclusion criteria for this research are summarized in Table 2.2 and any publications that met these criteria were included.

Table 2.2: Research inclusion/exclusion criteria

Inclusion Criteria	**Exclusion Criteria**
Field of medical and health education related to assessment	Medical and health education field related to program training and education
Assessment in MCQ examinations Undergraduate and postgraduate students	Clinical simulation examinations Kindergarten and pre-school ages
Using multimedia materials English language Other fields fulfilling the rest of inclusion criteria	High fidelity simulation and virtual reality Other languages Grey literature, abstracts, conference papers, unpublished articles, thesis dissertations

2.3.2.2 Bibliographic databases

A review of the medical, nursing and educational literature was conducted through the use of multiple engines. This search yielded only a few relevant studies in the medical literature. Only a few pieces of evidence were found linking MCQ testing with multimedia. This did not indicate that the literature wasn't available. Rather, it seemed

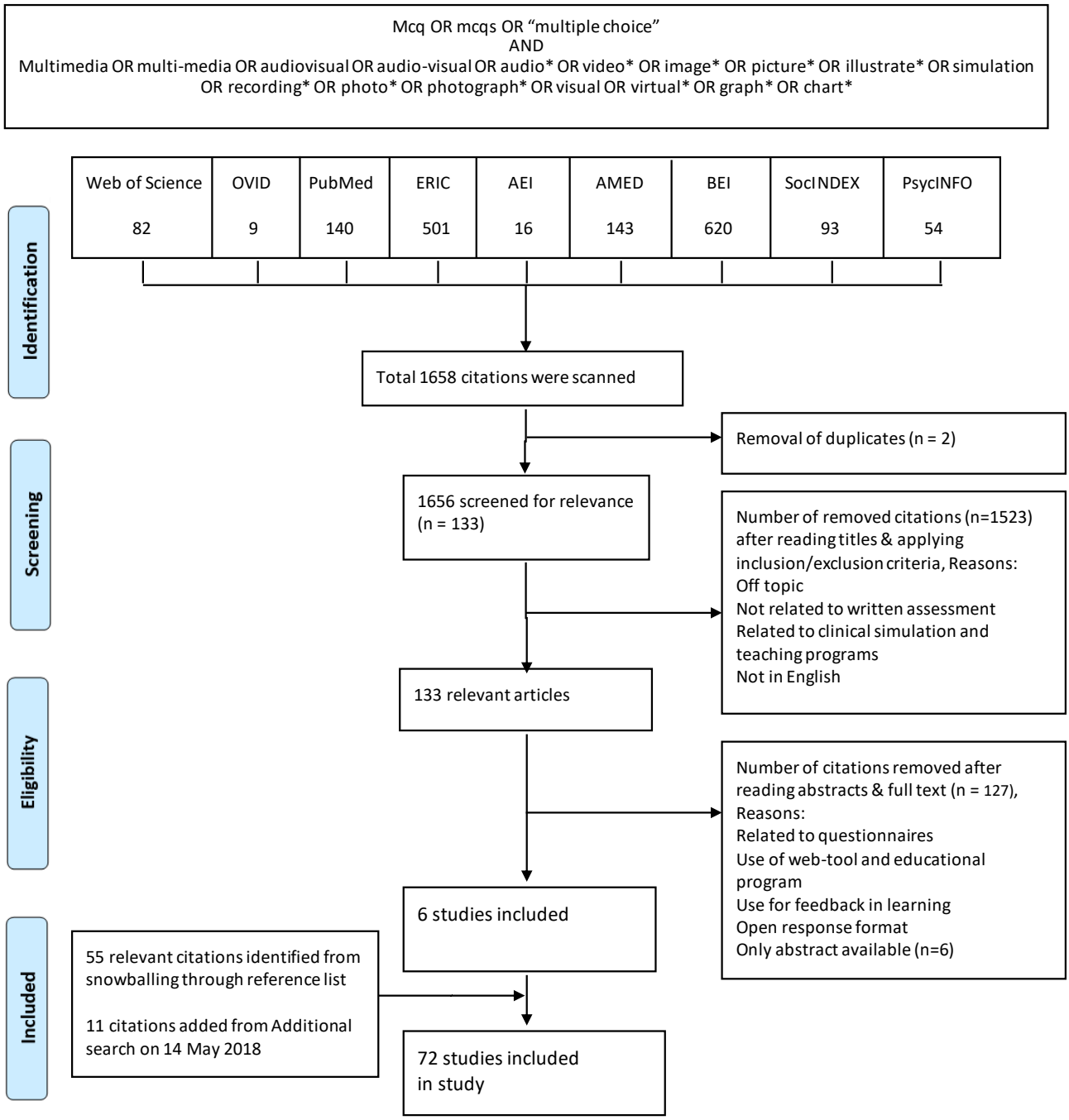
to be because of finding and matching the correct combination of the various possible keywords that were needed for the search. However, most of the relevant references were obtained through snowballing and reference tracking, browsing through the reference list and near-relevant papers (68). This was because the selected search terms were not always image, picture, or multimedia. In addition, after a further review of the reference list and extensive hunting for similar articles included, the related keywords to these articles were different for each with some having no relation to multimedia terms and synonyms, for example, keywords that were found were: problem-solving, educational assessment and educational measurement, computerized testing, digital testing, and technology-based assessment.

The following summarises the main findings of the literature review: The search engines and databases that were used in this study included searching MEDLINE (PubMed), OVID, EMBASE, CINAHL (Cumulative Index to Nursing and Allied Health Literature), ERIC (The Education Resource Information Centre), AMED, PsychINFO, BEI (British Educational Index), AEI (Australian Educational Index), SocINDEX, ProQuest, and Web of Science on October 31, 2014. The search of the literature was aided with the assistance of an experienced research librarian in the medical field. The search was last updated on May 14, 2018, with 11 additional articles added. Search terms and related keywords concerning multimedia in multiple-choice examinations were carried out using Boolean operators (AND, OR) (69 p.95-99) to either combine or exclude keywords and are present in Figure 2.1. The selected time frame for reviewing the published articles was from 1960 to 2018. Google and google scholar were also used in this research as a grey literature database.

2.3.3 Results of data extraction

Eligible studies that followed the criteria were included in this literature review and were evaluated by the researcher. Initially, abstracts and titles were assessed to meet the study aim. A total of 1658 articles were originally identified. However, after the titles were screened for relevancy, only six were related to the research topic and the rest were excluded for being off-topic. An additional 66 articles were identified and included, which were related to multimedia and assessment from the reference list reaching a total of 72 articles retrieved in full-text that were included in this review as shown in Figure 2.1. The reference manager Endnote X 8.2 was used to store and manage references and to identify any duplications of articles. Out of the 72 articles that were included in this study:

- 11 were related to low-fidelity simulation in assessment (written assessment)
- 14 were related to multimedia characteristics that would affect multimedia in assessment and
- 47 were related to multimedia in assessment, of which 11 of them were conducted on undergraduate or postgraduate medical education fields and had the same concept and method of this research study through exploring the effect of multimedia on item parameters (item difficulty and item discrimination) by comparing it to a text-matched item. These parameters are summarized below, and further details are found in Appendix 2.



Adapted from: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. & Van der Gijp, A., Ravesloot, C. J., Jarodzka, H., Van der Schaaf, et. Al (2017). *How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology*

Figure 2.1: Flow chart of included articles in the literature review of multimedia and assessment

2.3.3.1 Effect of multimedia on item difficulty (DIFF)

Out of these eleven studies, two showed that multimedia was more difficult than text-matched items, and only one of them was statistically significant (20, 70). Three studies demonstrated that there was no difference in difficulty level between multimedia and text items (45, 71, 72). However, when differences were noticeable, multimedia items were more difficult (45). Furthermore, there were another three studies that demonstrated that multimedia items were easier than text items (19, 40, 55), of which one was not-significant (40). Finally, the last three studies showed mixed results in difficulty levels (5, 49, 54), with one study explaining that when multimedia was difficult it was significant (5), and the other showed increased difficulty in cross-sectional images (54).

2.3.3.2 Effect of multimedia on item discrimination (DI)

Two out of the 11 studies showed that discrimination in multimedia items were more than text items (54, 70), one of which was non-significant (70) and the other one showed a significant difference in cross-sectional images (54). Seven studies demonstrated that there were no differences in discrimination levels between multimedia and text-matched items (19, 40, 45, 49, 55, 71, 72), with one demonstrating that when discrimination was noticeable, it was more increased in multimedia items (45). The final two studies showed that multimedia was less discriminating than text items (5, 20)

2.3.4 Discussion

It is apparent from the literature that the inclusion of multimedia to MCQs produces changes to item parameters whether it was a negative, positive, paradoxical or no effect at all. Multimedia items also seem to measure some aspect of skills and knowledge that are different from the print format, or they test different elements of the same concept. The conflicting results seem to demonstrate that multimedia items behave differently when interacting with an item and when interacting with the test taker. The type of media used, the number of items that were sampled in the studies, and the methods used to conduct them, also seem to have an effect. It should be noted that all these studies have limitations that include either having a small sample size, items that are not exactly matched, incomplete statistical reports, the use of a single method to obtain results, or no referral to validity evidence. Replicating similar studies will demonstrate if results from previous studies are consistent and provide a clearer understanding on the size of any effect. Further studies may be able to demonstrate under which conditions and context multimedia items become more difficult and discriminating than text, and/or vice versa. In addition, results from existing studies using the same methods should be reviewed and compared, to identify if there are certain factors that affect item parameters for text or multimedia.

2.3.4.1 Difficulty Level

When an illustrated item was more difficult, items were probably not clear and were of poor quality. These might have had an effect on students' ability to identify the relevant findings and, hence, their performance (55). It could also be due to the fact that students perceived some sense of responsibility towards these items and took the approach of one of less risky behaviour (72), which would reflect on how they would

act in their real setting. Another explanation why some multimedia items were perceived as more difficult was because they were more authentic and did not give out any hints, as did the text version (e.g., description of an audio-clip) (20). More so, these items were testing higher levels of cognition and so were perceived harder (45).

On the contrary, when an image was easier, some students of the middle and high ability levels weren't able to get it right, and when they had reached the correct answer it was for another reason. This could have been because of the image not being relevant in the first place, and, hence, not helpful, as well as the medium in which the information in the vignette was provided in (49, 55, 70). In addition, when an item was less direct in the description format, the multimedia item made it easier to understand; however, it was more difficult when it replaced textbook terminology, therefore, explaining why sometimes multimedia items were perceived easier or difficult according to the students (5).

In the study conducted by Lieberman et al. (2003), although their video-based questions had similar item analysis as the text items did, video items tended to not be significantly more difficult (70). While in the study carried out by Shea et al. (1992), the video format of the questions seemed to be slightly easier; however, the print format seemed to do just as well and, hence, if resources are available then the video format is advised to be used otherwise the time and cost of constructing video items is burdensome and challenging (19). This was also revealed in Hunt's study (1978), where visual data were found difficult to reproduce (55). Another study that used videos demonstrated that when item difficulty was noticeable video formats were generally more difficult (45).

2.3.4.2 Discrimination level

Discrimination of items will depend on the material itself and its relation to difficulty. If an item is too easy or too difficult no discrimination will be noticed.

Some studies demonstrated similar results in discrimination between the multimedia and the text items but also reported that when noticeable, even though it was non-significant, the multimedia items were more discriminating. This was suggested as a result of it testing higher cognitive processes (45). When multimedia items were perceived as difficult, discrimination was less (20). This could be because of the item being too difficult and requires one to have the knowledge of what he/she is seeing or hearing and, therefore, would either know the answer or not. While the opposite was seen in the Lieberman et al.'s study, where video formats were non-significantly more discriminating than the text items (70).

It seems that discrimination is difficult to interpret between multimedia and text items and is due to the fact that multimedia items seemed to measure some construct or element of the construct differently from what the text format measures (5, 19) and is as result of some type of context-dependent interaction that is taking place with the content (54).

2.3.4.3 Duration

In addition, it was noticed that time spent on multimedia items was longer than that of text items, which could be because of it being more difficult, it containing more information than the text, or perhaps examinees needed more time to identify the problem and were more careful about dealing with it as in real-life experience. It could

also simply be that higher fidelity items required more time in order to be answered (5, 20, 72)

2.3.4.4 Type of media

Some of the results in the studies suggest that the type of media (image, video) and case selected may render it easier or difficult than others (54). For example, in the study by Shea et al. (1992), some formats of cardiovascular studies, in precise the arteriogram cases, demonstrated different discrimination levels than the other cases in the text format, this seemed to suggest that certain skills are required to read motion studies (19). Students tend to behave the same towards multimedia items in a way they would benefit the patient and in a less risky manner (72). In the study by Holtzman (2009), the use of audio for CVS auscultation findings was perceived as more difficult and less discriminating compared to the descriptive version, which was easier to interpret (20). This could be that these types of media were not easily accessible and were more difficult to identify and differentiate than visual media. In addition, audio formats did not contain hints as would the description format. Therefore, only students who knew the diagnosis were able to answer it. Another example affecting item parameters was the spatial ability and cut of the image that was displayed. This was demonstrated in cross-sectional images that were perceived to be more difficult and discriminating as they seemed to require extra skills in being tested; while other types (schematic images) suggest a cueing effect, therefore, were more easier and less discriminating (54).

2.3.4.5 Students' preference

Whatever the effect that multimedia had on the test results, students preferred the illustrated and visual format of items over the text format for being more clinically

relevant (19, 45, 55, 71) and some preferred it to be included in their final certifying examination (55). This was because it relied on visual recognition that was used in clinical practice and almost all studies reflected on the importance of incorporating multimedia in examinations for that reason (55, 71). The use of computer-based testing to deliver these items made delivery easier and more practical (70).

2.3.5 Limitations

Although search terms were identified, its application was difficult and was not the same on each database, and so results would turn out negative in most searches. Therefore, each database may have had a different combination of search terms that were used after the original combination had failed. This may have led to some publications that may not have been identified and may have been missed. In addition, the wider grey literature PhD thesis, abstracts, and conference papers were not included in this search. However, the researcher feels that no major concepts have been missed for three reasons: 1) based on the reference list of relevant publications found, all articles seemed to be repeated and were included in this study; 2) other main studies that were found through snowballing reference lists were not cited in those previous publications; 3) a recent publication by Vorstenbosch et. al., (2013) explored the role of images in assessment and through conducting a search using MEDLINE, ERIC, Embase and PsycINFO and using different keywords and synonyms, ended up only finding one paper (Hunt, 1978) with no further hits after snowballing (54). In addition, the researcher feels that a major part of the literature review found was not covered in these articles and, therefore, were included here in the multimedia literature review that is presented in the next chapter.

2.4 Conclusion

The results from the literature suggests that there isn't enough evidence to refute or accept that multimedia items affect item analysis in a certain direction (e.g., more difficult or less difficult than text). More studies need to be conducted in different contexts and perhaps by using additional methods to gain a better understanding of the factors that may be affecting the parameters of the items (i.e., difficulty, discrimination, time). Validating the results of tests using multimedia items is important to ensure that the interpretations made are true and valid. This means that studies need to report what steps were taken to select and construct the items, the setting of the examination, the characteristics of the population chosen, the mode of computer-based testing that has been used, and technological factors involved, in order to gain a clearer understanding about the differences in performance of multimedia and text items.

The literature review has demonstrated that multimedia items are valuable, test at a higher cognitive level, and may measure a construct in a different way to text formats. Instructors should carefully select multimedia materials for examinations and closely examine their item analysis. Research is needed to understand how the combinations between these materials, the items, and test taker effect the performance of the assessment.

Chapter 3: Multimedia Literature Review

3.1 Introduction

In the following chapter, an introduction of general aspects of computer-based testing will be discussed followed by multimedia, the use of multimedia in learning and testing, leading towards the end of this chapter to the justification for this research project. A narrative review of the 72 articles from the literature review was undertaken, explaining and summarising the main findings in the literature and interpreting evidence related to the research question. A narrative review is a type of approach to synthesis and it integrates qualitative and quantitative evidence from published studies and other sources without any source of analytical purpose. They are commonly used in literature reviews (73).

3.2 Computer-based testing

Computer-based testing (CBT) refers to the use of a computer to deliver a test (74). It either provides all students with the exact set of questions as in the paper-based format, this is referred to as computer-based testing, or delivers a test where questions are based on the students' previous response and is referred to as computer-adaptive testing, (CAT), and therefore, in this format not all students receive the same set of questions (39, 74-77). Computerized testing is the general term for CBT, and it is sometimes referred to as e-assessment (78, 79), computer-aided assessment (79), computer-assisted assessment (66), and technology-based assessment (77, 80). In relation to this research study, the term will only refer to computer-based testing and not adaptive testing. Depending on the purpose of assessment, computer-based testing can be used at several points during a course or program as described by Smart C. and cited by Cantillon: 'prior to a course as a diagnostic assessment, during a course where students can self-assess their learning needs, at the end of a course

for feedback on performance (formative assessment) or at the end of a course where a pass/fail judgment is made for examinee qualification (summative assessment)' (66).

The process and concept of computer-based testing involve multiple stakeholders, various resources, and numerous challenges in order for it to function properly. Resources for CBT combines human resources such as (test administrators, supervisors, proctors or invigilators, examiners, item writers, IT technicians, programmers and supporters, as well as those involved in publishing and setting up the examination system), The other resources are logistical and technical in nature (81) and are concerned with setting up and providing a testing lab with computers and necessary equipment in addition to publishing policies and procedures for the exam process, testing lab, guidelines for examinee, examiners, proctors and test publishers that outline the responsibility of each stakeholder (81-83).

3.2.1 CBT and assessment in medical education

The use of computers for medical examinations have been documented since the 1960s to test knowledge and problem-solving skills by using MCQ and EMQs (1, 66). A great deal of research on CBT took place in the 1970s and operational administration of CBT begun in the 1980s (76). In 1986, with the growth of computerized testing, the American Psychological Association (APA) published guidelines for developers and users of computer-based testing on the development, use, and interpretation of computerized testing, emphasizing their responsibility to establish the validity of a computerized-test (APA, 1986, cited by Hao (39) and by Bugbee (74)). These guidelines complement the *Standards for Educational and Psychological Testing* (APA, 1985) by focusing more detail on computer testing, and emphasizing the developer's responsibility (74).

In China, a medical examination system with artificial intelligence had been used since 1989 to assess applicants for residencies in the medical field, including doctors, dentists, nurses, pharmaceutics, etc. (84)**.** In the 1990s, the United States Medical Licensing Examination (USMLE) started as a series of paper and pencil examinations that later was transformed, in 1999, into the computer-based format becoming the first generation of computer-based testing produced. The transformation took place to improve security issues that were associated with the paper-based format (7, 20, 85-88). With this new testing format, authenticity was introduced through computer-simulated patient format (known as the computer-based case simulation) that was under development by the National Board of Medical Examiners (NBME). It required examinees to select options for history taking and physical examinations and to manage simulated patients in recorded time (20, 66, 85, 88). In addition, all three USMLE steps have enriched the fidelity of their questions by incorporating the use of multimedia into their stems (20).

In other field areas, the Graduate Record Examinations (GRE) was available for examinees to take as computer-based tests in 1992, and as an adaptive form in 1993 (7, 74). In 1994, the nursing licensure examinations delivered their examinations only on computers (7, 74, 87). In the 1990s, medical educators introduced digital resources in their medical curricula, assessing medical students through CBT, as well as assessing postgraduate programs and licensure examinations in the UK, USA, and Europe (44, 66, 74, 89, 90). Since then, educational technology has played an important role in teaching and assessment (79) and a large number of organisations and associations have been taking the path to deliver their licensure and certification examinations through computers (44, 66, 74, 76).

In 2001, the International Test Commission (ITC) published the International Guidelines for Test Use that relates to competencies and skills that are needed by test users to carry out the testing process, as well as the knowledge needed to understand tests and their use (82). The document also covers implications for standards for test construction, documentation and test information (82). The ITC published guidelines that encourage the promotion of good practices that can be adapted for use across different cultures and languages to assure uniformity in test quality (82). In 2006, they published a comprehensive guideline called “The International Test Commission’s Guidelines on Computer-Based and Internet-Delivered Testing”. The guidelines are internationally agreed on guidelines of good practice for test-takers to use in educational, clinical, and organisational testing practice when conducting a CBT or internet examination (91). They are based on inputs from international countries, as well as various organisations including The British Psychological Society, the American Psychological Association, internet task force, the NBME, European Federation of Psychologists’ Associations and many more (91). In their guidelines, they highlight four main issues in testing (the technology used, quality of psychometrics, control over test administration and security/privacy issues) involving three main stakeholders: the test developers, users and publishers (91).

In the last three decades, modern computer technology has gained popularity becoming an unavoidable part of our lives (63, 74); CBT has been developing rapidly, and its use has decreased some of the financial burdens in test development with its user-friendly interface (39). With all these changes, individuals have become more computer-oriented and postgraduate and undergraduate medical students now have a higher degree of computer literacy than ever before (66). Computers have transformed our everyday lives through automated services, they have added value,

accuracy, and improvement to our educational and testing services, as well as made the delivery of multimedia testing both practical and feasible, deepening our understanding of the role of computer technology (39, 63, 92). However, there is still an urgent call for the delivery of high-quality computerized testing and even a greater reward for the role of computers in assessment and examinations (92, 93).

3.2.2 Benefits and challenges of CBT

Technology-based assessment is related to the use of computers and other electronic media, such as videos in educational settings, to assess an individual's progress (80). As explained by Schoech (2001), "a technology-enhanced assessment environment" provides characteristics for an examination. For the user, it allows test self-administration, a free navigations system that allows users to move from one part of the examination to another (94). For test developers, it allows for item and exam customization, implementation of innovative design formats (i.e., multimedia and virtual reality items) or interactive strategies (gaming and simulation), improvement of delivery techniques (using USBs, CD-ROM and secured net), automatic storing of results, immediate result calculation, and automated data interpretation (45, 66, 77, 94, 95).

Computer-based testing comes with its benefits that are not easily achievable with the conventional paper and pencil exams and this has been well documented in the literature (1, 7, 39, 44, 66, 80, 85, 94, 96). Computer testing enables diverse group administration through the availability of computer labs (66, 93, 94, 96). Unlike paper-based examinations, CBT allows for presenting cases and items more like those in the actual work settings by including multimedia and simulation and assessing decision-making skills and higher levels of cognitive construct. It provides an opportunity to

introduce computer simulation methodologies into testing (1, 7, 39, 45, 80, 85, 89, 97, 98). The use of images, videos, and sound into test items makes it more resembling to real-life tasks and enhances students' creativity, widens what can be tested (1, 44, 76, 79, 98), aids in assessing skills such as problem solving (39, 44, 66, 78, 89, 93, 96), and allows for better integration with the curricula (44, 78). It offers standardization in the testing environment minimizing errors involved in administration and score reporting, therefore, increasing test reliability (63, 78, 80, 85, 87, 96). It gives the opportunity to deliver tests to examinees almost on a daily basis in different locations and centres at their own convenient time (1, 7, 76, 80, 96). It is characterized by having a specified number of seats to it, allowing for a standardised comfortable environment and private testing area. Examination results and reports can be delivered accurately and immediately after finishing the test, which expedites decisions that need to be made (7, 39, 80, 96). CBT can improve cost and time and has the ability to maintain a large question bank that helps deliver and customize different examinations. This, in turn, further reduces security risks (1, 39, 77, 78, 96, 99). Feedback and information regarding items and test-taking behaviours can be collected and given instantly (39, 44, 74, 79-81, 93, 94). Computerized testing facilitates item analysis and can provide detailed information regarding the time spent on each item and items that have been reviewed by the examinee (39, 76, 97). Evaluation of psychometric data, item parameters, item bias, difficulty, discrimination and reliability can also be computerized promptly (39, 74, 80, 96). Finally, computerized testing can be customized and set to deliver items according to the examinee's responses, using computerized adaptive testing techniques (80).

Despite the many advantages of technology and computer-based testing, they still possess unanticipated "hidden" problems (63, 94) that may appear unexpectedly.

Because technology is continuously changing and the exam development process is often a multi-year process, computer-based exams need to be reviewed to ensure that they are not outdated when they have just been completed (94). The use of the internet allows for uploading, downloading, and streaming of videos into examinations. However, one drawback of the internet is sometimes the low quality of images or videos available and the inconsistency of the net, which can affect connection speed. It should be noted that delivering the exam via computer and, more importantly, via the internet has a risk of being compromised, particularly if someone taking the exam manages to capture a copy of it (94). All technical issues require the availability of an IT and technical support staff on-site and throughout the whole exam development process, to ensure that the examination continues to work if any problem arises and to ensure that no time is wasted to restart or have to change computers. These problems could be any of the following (78, 79, 93, 94, 96):

- Difference in layout,
- Hard disk freeze or crash, software bug,
- Hardware or software update or upgrade,
- Hardware or software disruption or failure,
- Change in the operating system or browser setting,
- Server maintenance disruption, network or power failure,
- Insufficient operating workstations and security breach Etc.

IT technicians with video production and programming skills are not regularly accessible and are one of the key links for the success of any CBT process (79, 94). Therefore, having the same technician available who is familiar with the project during the lengthy examination development process is not always possible (94) and the risk of remote staff or technical support being unavailable is likely and may occur (78).

A potential major issue in conducting computer-based examination is the difficulty in changing parts of the examination before the assigned date. As even minor adjustments can require one or several days of work. While in a paper-and-pencil exam, it is much easier to update the process of an exam (94). In addition, in the case that specialised test-developers were not readily available, there would be a risk of having a faulty examination (94). Another negative aspect of computerized testing, particularly for large-scale testing, is the number and location of centres that might be limited and if available would be costly (74). For those who are uncomfortable with computers, the experience of a computer-based examination could be stressful, particularly on top of the stress of an examination. This could be overcome by orienting the candidates to the computer interfaces or delivering a practice exam to allow users to feel comfortable before the actual exam date (80, 94, 100).

While the call for using digital technology is not new**,** most organisation apply the traditional assessment of the paper and pencil format, and most assessment using multimedia has been experimental. One reason could be due to the cost and the substantial additional workload it requires to get familiar with new software (94). In addition, much of the research in computer-based examination has been with populations of elementary school students and undergraduate students, and may not be generalizable to postgraduate high-stake testing (74).

3.2.3 Computer Vs. paper-based exams

Since the 1990s, there has been an increased interest in organisations and institutions to accept and convert their tests from paper-and-pencil (PNP) to computer-based testing (74, 101). Akdemir, Oguz (2008) and others have summarized results from the literature regarding students' performance on traditional paper-and-pencil format

compared with computer-based formats that yielded conflicting results (39, 63). Supporters of computerized testing stated that students' performance was enhanced, and stress was reduced while other studies reported lower test performance and anxiety among students who were unfamiliar with computers (63, 102). There were also other studies that found computerized testing to have no detrimental effects on students' performance (39, 63).

Bugbee 1996, reviewed a number of researches, reviews, studies, and guidelines about testing by computers and reached the following general conclusions: a) computers affect testing; b) special considerations should be taken when used in testing; c) computer-based tests can be equivalent to paper-based tests under stringent criteria; and d) it is necessary for test users to have a basic understanding of computers and knowledge of psychometric test properties (74). According to Liu, Papathanasiou, and Hao (2001), there are three factors that influence a student's perspective towards computer-based testing: their attitude towards computers, anxiety, and experience (80). Regarding anxiety, the authors demonstrated consistent views from the literature that computer anxiety does not seem to affect a student's performance. However, they did indicate from studies that there was a correlation between computer anxiety and low scores and that students who owned a computer were more prepared and had more experience, which alleviated some of their initial concerns regarding the format, hence lowering their anxiety levels and bettering their performance (80).

These studies, as well as others, have been showing various factors and conflicting results in regards to testing formats affecting test scores. With some aiding the advantage of CBT over PNP testing and others disadvantaging it due to the mode effect, the test-taking strategies (44, 74, 89, 96, 103), the increased anxiety level with

being unfamiliar with computer use, the variations in visual quality and computer screens in testing formats (86, 96), and the need for computer skills (96, 102).

The differences between computer-based and paper-based testing can include the way examinees respond to the questions, mode of exam review, progress overview, sense of control, test-taking strategies, the time limits, individual pacing, inclusion of multimedia, screen capacity, ease of response, test security, cost, as well as other factors that all need to be accounted for when trying to determine their equivalence (74, 77, 86, 96, 101, 102). Students who preferred CBT over paper and pencil tests listed the following reasons:

- Increase in motivation (104, 105)
- Reduction in testing time (104, 105)
- Ability to complete the exam at one's own pace (86).
- Image quality is better and independent from seating position (86)
- Rich interface allowing for a dynamic graphic presentation during testing (96)
- Direct feedback delivered to the examinee, as well as feedback to the computer for item analysis (96)
- Preferred the simpler way of answering the question through clicking instead of writing (39)
- Immediate feedback and score distribution were one of the most frequently cited reasons why students preferred the multimedia format according to Liu, Papathanasiou and Hao's study (80)

CBT also has the advantage of presenting clinical findings in a way that is consistent with what students face in their clinical setting, requiring them to interpret the multimedia stimulus rather than have the findings described for them (44). On the other

hand, PNP has the advantage of adding exam items at the last minute and creating a re-take examination is more feasible and acceptable than in a computer-based exam, which would involve creating additional multimedia items that would require an even more tiresome process if not available (94). The cost of CBT should not be underestimated and is higher than paper and pencil tests (66, 89, 101), particularly when testing for high-stakes examinations (66). The cost in CBT is mostly attributed to the use of test-delivering agencies (89) and also includes the cost for licensing of systems, installing software, technology malfunction, staff training and any hardware purchases (e.g., computers and OMR scanners) (77, 78). However, when the assessment life cycle is taken as a whole, computer-based tests are more cost-effective than paper-based tests (78).

Regardless of all these challenges, candidates' experience, as well as test-takers' acceptance worldwide and the perceived benefits of computerized tests, seem to outweigh those of paper-based tests that are less flexible (66, 74, 78). In addition, computer-based assessment is still considered a feasible way for test administration (105) that facilitates the delivery of a more valid assessment (66).

3.2.4 Simulation and multimedia in CBT

Simulations are viewed as exercises that are intended to mimic real-life conditions where students are placed in a situation where they are given the opportunity to reason through the presented clinical problem and decision-making process in a safe environment (46, 59). The aim of simulation is to reproduce and recreate patient scenarios in a realistic setting for assessment and feedback (47).

The use of simulation in computer-based testing can be technically complex due to its logistical problems. However, it may be carefully used as a testing strategy (92).

Simulation is used to assess physician's competencies in under and post-graduate medical assessment and is also used in the educational programs of multiple specialities such as anaesthesiology, surgery, obstetrics, emergency medicine (EM), paediatrics, and critical care (47, 48). Many of these specialities rely on a visual component to it because they are derived from the basic sciences of anatomy and pathology, which are mostly visually based (49, 50).

Simulation has also been introduced in high-stakes examinations (board certification and credentialing) (47) as a form of testing skills beyond that of which written examinations could elicit. It was designed to replicate some features and surroundings of the clinical environment (46, 67). Simulation can take many forms and can reveal many degrees of realism and authenticity on a scale that can range from a simple question on the screen, to computer-based cases, to real-life conditions. This, of course, depends on the purpose of the assessment and the skills that are intended to be measured (106). The computer-based case simulation (i.e., the virtual patient that the resident must care for), was first introduced by the USMLE in step 3 examination in 1999 (47, 67), as well as in the National Medical Licensing Examination in Italy (1, 90). Guagnano et al. introduced a new method called the MIPP project (Multimedia Integrated Pilot Project) that was administered in the Medical Licensing Examination in Italy. This was a single two-step examination comprising computer-based case simulations and multimedia that mainly assessed clinical knowledge and decision-making skills (step 1), and clinical encounter with standardised patients (step 2). Results demonstrated that this could improve postgraduates' performances (90). In 2003, the Royal College of Physicians and Surgeons of Canada (RCPSC) augmented their assessment of standardised patients by including simulation through the use of digitized cardiac auscultation videos. This came after they recognised that

standardised patients bare the limitation of not producing physical findings (47). It should be remembered that the use of virtual patients may be the most complex form of assessment in computer-based examinations and should be used carefully (1).

The MCQs are considered as a type of low-fidelity patient simulation format that is augmented with the patient's conditions and requires examinees to make clinical decisions (67). The observable outcomes from these questions (i.e., result of test scores) allow test developers in a way, to see into the examinees' abilities to apply their knowledge to a written description of their daily clinical encounters (67). Paper-and-pencil problem-solving items are also considered inexpensive readily available simulation formats that have been used for assessing residency training in different specialities, including the speciality of EM, in addition to the use of computer-based cases that use simple advanced life support (ACLS). The more expensive formats of simulations would include simulators, such as ultrasound simulators, anaesthesia simulators, and virtual reality cases (46). The patient-based MCQs (patient-management problems (PMPs) or written-management problems) as mentioned by Boulet and Swanson (67), have been used in medical education and assessment since the 1960s and later have been implemented in a number of specialty boards, as well as all USMLE step-examinations when it was first introduced in 1992 (67, 107). In 1999, the premium version of the computer-based case simulation (CCS) was introduced in USMLE Step 3 (67). The simulation presentation changed from step 1 (description of patient situation followed by questions), to step 2 (lengthier clinical cases with diagnoses and management questions), to step 3 (patient-physician encounter) (67). Up till 2004, multimedia was not extensively used in USMLE, and thereafter, USMLE has further increased the fidelity of their CCS through the integration of multimedia. All three steps utilize the advantage of computer-based

testing and enrich the clinical presentation through the use of multimedia in order to assess students' decision-making skills. These exam formats are viewed as a low-fidelity patient simulation method (67).

With the advances in technology and medical education, simulation technology had become user-friendly and presented an opportunity to improve assessment tools in high-stakes examinations by addressing certain inferences (e.g., physical abnormalities) and depicting physical finings realistically (e.g., skin lesions, breath sounds) that could not have been mimicked by simulated patients but could have been demonstrated through the use of multimedia (20, 39, 108). For example, the use of simulation technology has been used in previous studies where video clips were incorporated into neurology written examinations to improve its face validity. Studies demonstrated that this approach was feasible and reliable (70, 109). Other aspects that simulation can address are imaging results (e.g., CT stacks, US and X-ray), diagnostic studies, and the interactions among healthcare team members (20).

Although the literature is evidenced with the benefits of simulation use, it does have its limitation when being implemented. Some of the challenges being faced is a lack of faculty time and training, financial burden (expensive), and logistical matters (e.g., updating and maintaining equipment's, providing space for teaching and training, personnel hiring and budgeting) (47). Nonetheless, the use of multimedia aids in improving the fidelity of the item scenario (6).

3.3 Multimedia

The word multimedia, in the Oxford Dictionary, is described as 'using more than one medium of expression or communication; and in computer application: is incorporating audio and video, especially interactively' (110). The Association of

Medical Education in Europe (AMEE) Guide No.6 refers to the term “multimedia” to mean ‘courseware which integrates video, audio and graphical material with text and number operations’ (34). Visual displays, a form of multimedia, are tools for communication, teaching and learning, and are an important aspect of our daily lives. Audio is sound that can be presented in the form of a discussion between individuals (doctor-doctor or doctor-patient) or in the form of an audible clinical finding such as heart sounds, murmurs, breath sounds, wheezes or coughs.

Nowadays, they are not only present in books but are also present with us in our homes, at work in our computers, at school in our learning materials and in our everyday hand-held devices (111-115). Not to mention, the World Wide Web which had arrived at our desktop computers in the 1990s and was rapidly being developed, that with it, came a great deal of multimedia information (115, 116). Today, we are exposed to more images, graphs, and videos than we were a few decades ago (111) and the amount of information that one can acquire through visual mediums has dramatically increased. With this has arisen the opportunity and ability for researchers to understand and evaluate the aspect of “visual representation” in the field of education (114). However, it seems that only a few studies have explored the analysis of documents containing images and text in assessment, and within multimedia, the focus was on images and videos and, to a lesser extent, on audio-media types (115).

3.3.1 Authentic assessment

One of the benefits of computer-based testing is that it can be used to deliver “authentic” assessments through items that are able to simulate real-life scenarios. Authentic assessment requires examinees’ to demonstrate competencies (knowledge, skills and attitudes) in situations that resemble real-life context in a similar way they

would do in the real clinical setting (criterion situation) (66, 117). Authenticity is subjective and changes according to personal perception, age, educational level and professional experience (Honebein et al. cited by Gulikers (117)). However, the more an assessment is authentic and resembles the notion of the real professional practice, the more its predictive validity increases (117).

Multimedia items, as well as being a type of simulation item, are also a type of innovative item (45, 95, 97). Innovative items as described by Parshall et al. (2014) 'are those items in a CBT that make use of features and functions of the computer to do things not easily done in traditional paper-and-pencil assessments' (95, 118 p.1). They involve degrees of interactions and performances by the examinees and include audio, video, or animation even if the item stem itself is in the conventional multiple-choice format (99). An innovative item can be complex, requiring an examinee's response and interaction with a virtual item multiple times, or can be as simple as an MCQ that requires viewing a set of images to answer a question (e.g., view images through a microscope) (59, 97). Other terms for innovative items that may be found in the literature include: 'computer-based items, technology-enhanced items (TEIs), technology-enabled items, technology-enhanced innovative item, perceptually motivated item types (PMIT), and innovative computerized test items' (5, 45, 59, 77, 95, 97-99).

In the field of medical education, one of the essential skills of competency for most if not all medical speciality graduates, is the interpretation of visuals and perhaps, to a lesser extent, audio data (55). It is almost inevitable to have a medical trainee graduate without having these skills. However, in the absence of a clinical setting where an examiner and examinee are face-to-face, it is difficult to test an examinee's ability of visual recognition. OSCE or other objective clinical examination can only test a small

set of these skills. This and other factors have stimulated the growth and use of multimedia (visual and audio) materials to be included in MCQs and other written examination formats (56). Like MCQs, the view on the potential of multimedia application is conflicting. For some, it is seen as a new trend in educational learning and teaching environment that possesses high expectations but lacks in delivery. For others, it is seen as a ground-breaking method in the conventional educational world (34). Nonetheless, innovation in technology allowed for innovation in multimedia use, developing new formats for test and item construction (94). Multimedia possesses interesting possibilities and allows for the opportunity to better measure higher cognitive skill and related aspects that were previously not available and were omitted from the conventional format (7). It also provides a means for understanding students' ability by identifying relevant from irrelevant information (119).

Multimedia can be implemented in a sophisticated complex way as in assessment, having a user experience being drawn into a virtual real-life environment. It can also be as simple as text-based lectures linked to databases of images (34). And like with any new element in technology, it is tempting to immediately dive into its application without fully understanding its use, implications, and consequences. In order to have a meaningful experience and interpretive results when using multimedia, we need to justify its proposed use, ground it into theories, and apply appropriate validity frameworks. Gathering such evidence will guide on how, when, and why to use multimedia in large-scale examinations (7). The multimedia format is generally accepted. It offers new exam possibilities and is characterized by being dynamic and flexible when delivering items. As it allows for any combination of images, sound, graphics, videos or animations to be implemented (39, 94), it is more interactive and perceived by its users as being realistic (94). In addition, the availability of ready-to-

buy software, as well as the opportunity to have customisable ones are not uncommon nowadays, and their associated databases are usually equipped with the capacity to store large amounts of data such as images and videos (94).

3.3.2 Potentials and drawbacks of multimedia use

The use of images, photographs, pictures and tables have been included in examination questions to test students' abilities in interpreting them (120). However, research in this area has been rare in comparison to the research demonstrating the use of visual resources in instructional design in teaching and learning (38, 120). Results demonstrate that learning and retention are better with the use of pictures and that students remembered better when receiving an illustrated text (120, 121), as these items tend to align more closely to the curriculum (97). In testing, they tend to test a greater depth of students' knowledge measuring a broader range of skills and higher-order thinking (59, 97, 99). Illustrations, graphics, and videos tend to have a motivational role (39, 99, 120, 122), can provide more information than text, help clarify and interpret text content that is difficult to understand, improve complex and dynamic information, reduce the length of time spent on a question, and simplify information that may have an abstract or complex concept (39, 97, 99, 120). Furthermore, innovative items help reduce guessing, reading load and demands on working memory all of which reduce construct irrelevance variance and allow for a more valid measurement (97, 99, 118). All these are equally applicable to exam questions and are justifiable reasons to use different visual resources instead of textual descriptions in examinations (120, 123). The use of visual resources in examinations could also have the effect of making a question more interesting or seem less daunting to students when they are in a stressful situation and time (39, 120). It also allows instructors to write valid and creative questions to evaluate students' understanding of

important concepts that are related to real-life practising (122). In the learning environment, videos have the ability to situate students in a realistic clinical scenario that promotes authentic learning, stimulates curiosity, and engages their attention (124); and in the testing environment, innovative items are more authentic and engage students more (95, 97, 99).

The use of multimedia and visual representation is favoured for its capability of representing information in a novel way, obtaining knowledge for learners who are unable to get the information from text alone, capturing their attention and mainitaining their motivation (114). An even more important quality is that it enhances information retention of the associated text, which improves problem-solving and integration of new and prior knowledge together (114). Visual representation has the ability to provide information and phenomena that might be unavailable (i.e., too abstract, invisible, small, large, fast or slow) to the naked eye and transform it into a visible illustrated or dynamic representation that displays the information in an appropriate way. In simple, visual representation has the ability to transform complex data and phenomena that are difficult to describe or see into one that is understandable (114).

In the clinical simulation test (CST) format, nursing examinees reported that they found the audio-visual (AV) format to be more enjoyable, more satisfactory interaction with the patients, more realistic, felt more involved with CST patients and felt that they were able to identify problems more quickly. Nursing examinees who had the audio-visual clinical simulation format significantly took longer to complete the exam and were found to commit less risky and inappropriate actions than those who took the paper-and-pencil version (72). In another study, undergraduate students particularly liked the visual interactive aspect of the multimedia examination because it was more engaging, there was no need for rote memorization, and it resembled more of a problem-solving

case where more information was revealed when answering correctly (80). They also felt that the visuals helped them recall information and, hence, spend more time understanding the concept and less time trying to memorize the details. Students were able to relate to the relevance of multimedia to their actual practice giving examples of using a 3D model during class and then in the examination (80). Teachers also felt that the use of multimedia made it more possible to make the questions more interactive, dynamic, and authentic, enabling students to engage and analyse information. These teachers were enthusiastic about using multimedia in their teaching (80).

The most commonly cited reason for using technology-enhanced items is their ability to conceptualize the concept, measure higher-order cognitive skills (such as reasoning, synthesis, and evaluation) and problem-solving skills that are not easily assessed by MCQs (39, 97). There is evidence that demonstrates that the use of multimedia (particularly the use of scientific diagrams) encourages the process of logical formal reasoning, which is what examiners seek to assess in their tests (120).

And while there are new potentials with multimedia use, there are also modifications, delivery complications and other technical aspects that come with it (94). Opposing views regarding the beneficial use of visual resources demonstrate that the effect of images is somewhat unpredictable and caution is required (54, 120) with some studies showing that including images in instructional texts had little or no effect and suggested that the choice of image selection and appropriateness was what was considered important (120).

One of the most important negative aspects of multimedia materials was its quality. Where as a result of poor image quality, examinees were unable to identify and

interpret the materials presented to them (55) and were not motivated or interested to continue reading items accompanied by poor diagrams (120). In addition, an examinee's location in the testing room may disadvantage some in regards to some having a better view of the materials on the screen and better sound quality than others (95).

Another explanation for the failure of visual resources can be attributed to the students' learning styles, as not all students are adequately able to process images equally. Some claim that the use of images with text may have a harmful effect on students' attention, as attention had to be split between two forms of information, which then need to be integrated (120). Some disliked the format because it involved them emotionally with the patient's case and affected their test performance (72) and others felt anxious about using the multimedia format (80).

Multimedia or (visual resources) play an important role in developing a student's mental model. While not all observers will use the image in the same way (120), some students observe themselves as being visual learners and appreciate the use of visual resources in questions (122). Therefore, one of the main risks that visual resources can impose in the context of an exam is diverting the students' attention from the text to the illustration, distracting them and leading them to construct a mental model (or representation) to the text of a question that does not fit the intended meaning of the question (39, 120). It should be considered that the inclusion of an unnecessary, incomplete, or uninformative image may lead students to pay too much attention to it at the expense of what is relevant in the text (120).

Regarding the technical aspect, some students experienced technical difficulty when computers sometimes crashed, froze, or the software did not work properly (80), a

number expressed not feeling comfortable with dealing with computers and digital media, and a few mentioned that reading from a screen was tiring (80). Another example of technical difficulty was related to multimedia size and use. High-quality videos that are longer than 30 seconds are usually large in size, bulky to edit, and problematic to copy (94). Even with large video files (30-80 MB), their display on screen are usually not more than one-fourth the size of a computer screen and require a fast computer to capture and play the video (94). In some cases, it was suggested to segment the video clips into smaller sections (94). However, this action may affect the quality of materials. An additional way is the use of a portable mode of delivery (e.g., CD-ROM, USB or hard-drive), but nonetheless, this would also disturb the security of item delivery as it is more subjected to the chance of being copied and distributed (94).

Perhaps the most negative downside of multimedia examinations is that it is very hard to create video or image-based items during "re-take" examinations (122). It is difficult to develop a large multimedia bank, particularly a video-clip collection, that is different from those used in teaching (122). This would require the efforts of writers in providing a steady stream of images and videos and collaborating with libraries to acquire relevant clips (122). Video-based items also have the disadvantage of the need for reviewing the clip more than once, which can reduce the time and self-pace nature of the exam and may prevent students from being able to review other clips more than once (122).

Finally, multimedia materials can be more expensive to develop, their performance characteristics and impact on test-takers are not fully understood due to the variety of items, and they may introduce construct-irrelevant variance (99). The drawbacks of cost, difficulty of establishing reliable measures of performance, and logistical

complexity can be overcome by the use of computer-based testing and modern psychometric methods (39, 53).

It has to be kept in mind, that most of these researches were conducted in laboratory settings and were more focused on text supported by images. Therefore, their conclusions may not be generalizable to the educational and assessment settings. Nevertheless, with all these drawbacks, it is still believed that the benefits of multimedia outweigh its problems (122), that the use of multimedia is very likely to have an effect on the learning outcome (54), and would be expected to have a similar effect on assessment.

3.3.3 History of multimedia in testing

The use of multimedia and innovative items has shown to measure a different aspect of reasoning from standard written items (5, 7). However, the use of multimedia has uncommonly been used in large-scale testing because of feasibility issues (7). The following are examples of earlier computer-based test research in different fields outside that of medical education that incorporated multimedia into their examinations.

Multimedia in assessment is not an original idea neither in medicine nor in the wider literature. Using multimedia, particularly the use of film in large-scale testing can be dated back to the 1960s, where it was used by the US Army Air Force Aviation Psychology Program during World War II to test for decision making (Siebert and Snow, 1965, cited by Bennett (7)). The use of audio clips was usually restricted and primarily used in the areas of language and music testing. Examinees would be given a headphone and would listen to a lecture or spoken passage after which they would answer a series of MCQs (7, 95). Later on, audio clips were used in other areas such as the English proficiency test for non-native speakers and the advanced placement

language tests (7, 95). Bennett et al. (1997) discussed the application of historical radio spots in history, and the use of heart sounds in the sciences field (7) that supplemented and reinforced what the examinee was reading (95). Multimedia has also been used in history tests by using what is called non-text-based sources, such as paintings, architecture, maps, and cartoons to aid in providing information about the past (Beschloss, 1997; May & Zelikow, 1997, cited by Bennett (7)). As historians became more interested in these types of formats (such as films, radio and tape), history teachers started to recognize their importance and started using document-based questions, which incorporated multimedia materials (e.g., using maps and capture frames with a question about the depicted events) (7). In the speciality of physical education, which is a movement-based discipline, examinations incorporated a 1-2 minute video recordings of children's movement forms to assess a physical educator's knowledge and ability to analyse movements. This type of display would be restricted in a paper-based medium as the key construct wouldn't be able to be demonstrated for assessment (7). In the field of science, laboratory sciences and health allied sciences assessment contained items that used static electrocardiogram strips, dynamic traces from heart monitor displays, as well as heart sounds for examinees to diagnose patients' conditions. This would have not been feasible in a paper-and-pencil test either (7). Examples are also mentioned in the field of interior design, where students who took the computerized version of a test (with animation) had average scores higher than those who took the paper version, which was similar in content and sequence. The author suggested that the assessment with animation helped students improve their visualisation skills (125).

Research conducted outside the field of medical education demonstrates conflicting results regarding the use of animations (or illustrated phenomena) when compared

with static images. This might suggest that different media types affect the cognitive load differently (38). Berends and van Lieshout conducted a study in 2009 on the effect of illustrations on the accuracy and speed of solving arithmetic word problems (126). The results revealed that illustrations had a hindering effect on performance (making it slow and less accurate) when information was irrelevant, redundant, or interacting with the item. Furthermore, adding illustrations to word problems did not necessarily lead to improvement in performance (126). In 2010, the Department of Anthropology at the University of Texas in Austin conducted a mixed-methods research study to investigate the difficulty levels of multiple-choice multimedia test items and if they aided students to answer questions correctly (39). The findings disclosed that multimedia format had significant relationships with item difficulty levels. When items were easy (item difficulty index between 80-100%), the text-only items were easier to answer than the multimedia items. On the other hand, when test items were of medium difficulty (item difficulty between 51-80%), the formats competed, but most of the time multimedia items were easier to answer. When items were in the higher range of difficulty levels (item difficulty between 21-50%) the multimedia items were easier to answer than the text items. Finally, at the highest difficulty levels (Item difficulty <20%), the text-only items were easier to answer although the differences were not big (39). Students' behaviour and perceptions regarding the multimedia exam were collected through the think-aloud process and individual interviews (39). Participants felt that multimedia items helped them understand the content more and most had positive attitudes towards it even though they felt that some were unnecessary or unhelpful (39).

Another study looked at integrating animated traffic scenarios and whether it would enhance elements of the competence of driving in an MCQ German driving test. Two

versions were created, a static MCQ exam with still pictures of traffic scenarios and another dynamic version that contained animations of similar situations. Results were positive, suggesting that animations had an effect on the quality aspect of driving assessment; however, further research needed to be conducted due to the low criterion validity of the test (127). The following section further delineates examples on multimedia in written examination, particularly on MCQs and their implications on test parameters.

3.3.4 The history of multimedia with MCQ written examinations

Future generation of tests need to include tasks like those encountered in the real world in order for them to measure problem-solving, decision-making, and other cognitive constructs successfully. This could be relevant through the use of multimedia, by incorporating images, video, audio, animations, and dynamic stimuli. And with the dawn of computer-based testing, this made the inclusion of such a task more possible (7). Although computer-based testing has existed for more than decades, the use of multimedia has been limited and most CBT has been restricted to certain types of tests (80).

Many high-stake examinations consist of a narration of a clinical situation through the use of text in combinations with static-image-based MCQs (5). The use of MCQs in examinations has mostly been used to measure recall of factual knowledge and interpretation of examinees' abilities depending on the purpose and stake of the examination. Although the issue of illustrated MCQs has been raised by Hunt (55) in 1978 (almost 40 years ago), the literature is still scarce with research conducted to explain the effect of introducing multimedia materials (video, audio, or both) into these types of questions. Moreover, little has been done to critically assess them.

As shown in the table in Appendix 2, the aim of Hunt's study that took place in 1978, was to construct tests that moved from the cognitive domain of recall to interpretation, and in order for that to occur, written descriptions in test questions were replaced with radiographic images to allow examinees to interpret them as they would do so in their clinical settings (55). In Hunt's research, illustrated items (radiological images) were more difficult than text items because students had to interpret the images, which was considered a complex "extra task" when compared to the control group that had the description of clinical findings (55).

About a decade later, in 1987, Buzzard, Bandaranayake, and Harvey (1987) wrote a paper on 'How to Produce Visual Material for Multiple Choice Examinations' (56). In their paper, they discussed the print format of a postgraduate examination and what techniques were used to produce the print format of radiographs, CT scans, histological and other materials. They provided guidelines for producing high-quality prints with indicators, and stressed the importance of working closely with a photographer to produce the images. This was to ensure fairness for candidates and that they would be able to answer the items correctly (56). Another study in 1991 was carried out by the same authors to determine if MCQs based on visual materials were more difficult and more discriminating than equivalent verbal MCQs that had the same information and another set of verbal MCQs that had different information but was in the same content area (71). Their study demonstrated no significant differences between item formats neither in difficulty nor in discrimination.

Studies have demonstrated that the traditional MCQ format is fairly flexible and is able to measure skills beyond knowledge (19, 128). While other studies argued that MCQ tests are not flexible enough to capture the multifaceted and critical components of decision making and problem-making skills that are required in the health field (3, 19,

72, 129). MCQs have been scrutinized for having a cueing effect on patient problems through the presentation of questions with their answer options. A research study investigated the relationship between performance on MCQs (cued format) and free-response formats yielded conflicting and controversial results. The cued format may have had a tendency to overestimate abilities in medical simulation and students who took this format were found to have had higher success in selecting the correct diagnoses and, therefore, a higher estimate of performance (72).

In the following years, significant efforts in medical education have surfaced to develop innovative testing methods designed to present candidates with a more authentic experience and more real-life and problem-solving challenges. Examples of such methods are the use of standardised patients, OSCE, video, computer-simulated presentations, and computer-based testing that were used in conjunction with the typical paper-and-pencil MCQs or as alternatives to them (19, 72). However, much controversy has resulted regarding the cost, efficacy, and psychometric soundness of such methods (72). Shea et al. (1992) conducted a series of correlations of MCQs with cardiovascular motion studies and found that performances on motion studies (video presentations) and MCQ items with similar but not identical content had a low to moderate correlation in scores. However, the print and video subsets had a high correlation (see Appendix 2). This suggested that MCQs and motion studies measured different knowledge and skills but print and video subsets were equivalent (19, 72). This suggested that the higher-fidelity video-based items tested something different than lower-fidelity questions (19).

The Cardiovascular Disease Board of the American Board of Internal Medicine believed that cardiologists should be able to have the skills to interpret motion studies (e.g., include echocardiograms, ventriculograms, and arteriograms) in their practice

and included these kinds of studies in their examination blueprints (19). The decision left to be made was whether the format of these studies would be presented in a print format with still photos, or in actual motion format. In 1989, a study was conducted to compare video and print presentation of motion-study cases to determine whether both formats were equal for use on the certification examination (see Appendix 2). The study demonstrated the equivalence of the video and print format motion studies. It also supported the use of the print format in national examinations, reserving the use of video format locally when proper resources were available. Another study reached the same outcome that computerized tests were equivalent to written tests, and that differences in testing times were due to the students being unfamiliar with the computer format. However, students clearly preferred the computer administrated test (80).

In 1993, and after field testing that was conducted in 1989, the American Board of Internal Medicine included a new test-item format in their certification examination of the cardiovascular disease speciality. This format consisted of selected still-frame photographs (echocardiograms, ventriculograms, and angiograms). The study reported this first administration of 43 motion study cases in the examination and emphasized on its validity evidence. This motion study was included as a separate section because it was seen as vital for a cardiologist to be able to interpret them in their practice (130). In their study, validity evidence was provided through thorough explanation of the training, as well as in selecting and reviewing the content of the examination to be meaningful and important for clinical competency. Evidence was also provided through the description of the scoring procedure, providing statistical evidence and correlations to other scores in the examination and to program directors' ratings on examinees' performance in the clinical setting (130). The study

demonstrated successful implementation of the still-frame motion-study format and showed evidence that the use of still-frame motion study cases provided measurement of the intellectual competence in the cardiovascular disease certification examination. It also suggested that simulation of complex clinical behaviour can be demonstrated in other contexts instead of the more costly ones (130).

In 1994 (72), an article was published that tried to study the effect of interactive AV enhancement versus text format on the examinees' decision-making process, and problem-solving skill in a nursing competence computerized CST. Results showed that no reliability was lost by a change in format (i.e., internal consistency reliabilities for AV and text format were similar) (72). There were no statistically significant differences found in the ability estimates between results on the AV and text formats, indicating that both groups behaved similarly when it came to their actions towards benefiting the patients (72). In 1995, the AMEE published their 6th guide titled: 'AMEE Medical Education Guide No. 6. Evaluating multimedia applications for medical education' (34). Although their article tackled the use of multimedia in the perspective of the teaching environment, their methods and philosophies did not go as far as to how they could be used in the assessment field (34). The main question that educators and assessors would have liked to know as proposed in the article was: 'does the application of multimedia in any setting make us more conscious and critical of our own practice?' (34). Is there a transfer of experience from an artificial environment to an actual working environment, and can these be translated to testing materials? Do the behaviours of health professionals alter or improve when having been subjected to the multimedia experience? The conclusion of the article was that we still do not know to what extent transfer occurs from a simulated multimedia experience to the way professionals work in their clinical settings. It is for sure that transfer does occur to the

extent to which it would depend on the authenticity and lifelike of the simulation that has been taken. Research in this area of transfer is still needed in medical education (34).

Assessment plays an important role in determining the learning behaviour, as students will style the way they study in response to the exam content and format (54). As mentioned earlier, assessment needs to be at the same stage and level as teaching methods. Vorstenbosch et al. (2013) stated that 'Modernized teaching methods demand modernized assessment methods' (54 p.1), and at the start of the 21st century, there was much interest in the integration of multimedia into assessment, particularly at the college level. This was mostly due to the current trend of integrating multimedia into teaching by faculty and the availability to do so even with those possessing limited computer knowledge. From here, innovation in assessment had to be aligned with those used in teaching using computer technology and web-based education (47, 80, 131). For example, the use of video in health professions education has increased in the past decade, with Stanford University School of Medicine incorporating short videos into their curriculum(124). In addition, well-designed videos and educational materials are available on the web, where medical students can attend to their individual learning needs and review the content at their own pace through digital format (124). Therefore, it is important that knowledge regarding different assessment methods do not fall behind multimedia teaching methods (54).

Using videos of patient clips in computer-based testing evaluates examinees' skills in interpreting physical examination findings. However, the characteristics of video-based question psychometrics are unknown (70). A study conducted by Lieberman et al. (2003) compared parallel test questions of video clips and text descriptions of abnormal neurological findings (70). The authors developed a computer-based

examination for fourth-year medical students to assess their competency in neurological disorders through a test that contained both text- and video-based vignettes. This was conducted to further understand how using higher fidelity multimedia patient video clips may affect an MCQ's psychometric properties (70). Preliminarily analysis did not reveal one medium to be superior over the other, but it seemed that video-based questions increase the face validity of test questions and proved to be more discriminating than questions that were text-based (70).

Most items presented in written examination do not resemble real-life applications of knowledge and skills to the point that evidence of inferences from the results regarding examinees' skills is considered weak (70), for example, trying to interpret physical examination findings in written tests that are considered limited in its validity by the assessment's low fidelity (70). The implementation of novel media, such as these video-clips with MCQs aided by the use of CBT has made it possible to increase the authenticity and fidelity of medical licensure examinations by making them more similar to the actual tasks that are required in the clinical setting (70). Brooks et al., (2000) as cited in Lieberman et al. (70) reported that the use of images, such as patient photographs with a brief clinical history, improves the diagnostic accuracy of both novices (medical students) and experts (academic internists). The addition of image description and features further enhanced the diagnostic accuracy of both groups (70). And, as expected, experts scored higher than novices; however, they scored greater when an image was present. This provided validity evidence for the use of visual multimedia to assess diagnostic ability (70). Further evidence was needed for greater validity of video versus text-based questions (70).

In an article titled, 'Can a picture ruin a thousand word? The effects of visual resources in exam questions' the authors constructed experimental test papers for secondary

schools, that included two versions of six questions with graphical elements (120). The aim of their study was to investigate the visual resources (e.g., pictures, photographs, tables, diagrams) used in the examination to understand whether refined changes to these features affected how students understood the questions (120). After the tests, 27 pairs of students were interviewed. They concluded that visual resources should be appropriately designed and that careful consideration should be taken when incorporating them into examinations. If instructions were not carefully followed, then textual information might be better to use. Educators should also take into account individual differences among students when dealing with visual representations (114, 120). In 2007, the USMLE introduced multimedia-based presentations of cardiac auscultation findings into Step 1 and Step 2 Clinical Knowledge (CK) and compared it to the text format for investigating the impact of multimedia on item characteristics (20). Since their introduction of multimedia in all three USMLE steps in 1999, little has been known about the influence of their use on difficulty, discrimination and response time (20). As shown in the table in Appendix 2, multimedia items were perceived to be more difficult and less discriminating than their text version. One reason could have been that examinees were aware that auscultation skills were being tested, and item parameters would narrow if examinees spend more time studying for it. Another reason could be that the study was focused only on heart sound interpretation and the results could not be generalized to other types of multimedia or multimedia presentations of other findings (20).

Hertenstein and Wayand published an article in 2008 providing empirical evidence for the use of video-based test questions (122). They stated that while many psychology instructors used videotaped examples in their courses, only a few actually used them in their examinations, and proposed that they should be included (122). The authors

offered guidelines for video-based questions use, as well as benefits, drawbacks, and utility that could be applied to a variety of disciplines and levels (122). Results of their study revealed that students preferred video questions more than MCQs, that it deepened their level of understanding and allowed for a fair assessment of their knowledge and of real-life behaviour when compared to the MCQ items (122). The use of video-based questions allows for evaluating students' understanding of concepts and procedures in a novel context, as knowledge must be applied outside its original context in which students initially learned it from (122). However, the results of the study were unclear whether students' performance on video-based questions could be translated into their abilities outside the classroom (122) and would need further empirical research.

In the years 2008-2009, the Osteopathic Medical Licensing Examination conducted a study with multimedia text-matched items to determine if multimedia items were able to test medical knowledge and skills and how to produce effective multimedia items (5). They concluded from their research that multimedia items were capable of measuring different elements of the construct when compared with their text-matched items and that it is possible to develop multimedia items with reasonable psychometric properties (5). This study was different from previous studies conducted that focused on group-level comparison between paired items. Although this is necessary, the authors of this study took a detailed item analysis approach of comparing individual item pairs, as well as content analysis of the multimedia items in order to get a better insight on the behaviour of multimedia items (5). The results ended up demonstrating different statistical properties that were significant for some multimedia items. It suggested that the difference in the behaviour of multimedia items were due to it measuring something different (5). The study also demonstrated that the difficulty in

the item level (being more easy or difficult) was more likely to be as a result of the amount of information that was either added or removed by the given multimedia content in comparison to its text-matched item (5).

Results from a study on the use of innovative questions to improve assessment of nursing practice demonstrated that these items were more difficult and better discriminating than their counterpart in the text format. Students felt these items were more representative of their actual work performance and that video items required more cognitive skills and was perceived as more authentic than the text-matched MCQ items (45). In the few instances where the text-matched questions were perceived to be more difficult, it was thought to be related to the content error or the examinees' lack of familiarity with the content. Evidence from the results seems to support the development of such items for the NCLEX program (national licensure examinations for nurses) by the National Council of State Board of Nursing (45).

One of the other specialists that count on images in their assessments and examinations are anatomist (54). A study conducted in 2013 by Vorstenbosch et al. (54) aimed at investigating the effect of different types of images on item parameters (difficulty and discrimination) in written assessment. This study can be viewed as a contrast to Hunt's (1978) study. In Vorstenbosch et al.'s. study, their design focused on the response format (selecting the answer from a labelled image and an answer list) in written assessment and only included the use of images (54). In their study, they looked at the influence of images as a response format on item parameters. Their research demonstrated variable effects from images suggesting that a context-dependent interaction was taking place with the item content. Images influenced item difficulty and to a lesser extent discrimination. In addition, their study revealed that cross-sectional images seemed to test an extra skill (54).

Furthermore, another study was carried out on first-year medical students to try and gain insight into the effect of images on the validity of test items. In this study, however, students were given EMQs on gross anatomy items that were either combined with labelled images or an answer list, and their cognitive processes were examined through the think-aloud process (50). Results demonstrated that students often used hints or cues from an image or answer list, as well as knowledge, followed by visualizing, verbal reasoning, and eliminating to reach the correct answer. The study concluded that EMQs with and without images yielded different results that suggested different cognitive processes were taking place, hence measuring different skills that made them valid for different test purposes. (50). This indicated that the response format influenced the validity of the stimulus (50). In the speciality of histology, a recent study was conducted comparing illustrated and text items in order to demonstrate whether the inclusion of images within the stem (stimulus format) of a histology MCQ could have a consistent or predictable influence on item psychometric properties. Overall, results showed that there was no influence on item parameters (40). A more recent study in 2017 took a retrospective analysis of 15 MCQs in the speciality of anatomy. A comparison between text-based items and items with images were reviewed to see if the presence of images affected item analysis. Results showed that there were some differences in item difficulty but they were not consistent with either text or image items. The conclusion was, that images do not significantly alter item statistics, and suggest that if images were to be added in assessment, then instructors should select accurate and appropriate images to make sure that no adverse effect comes out from it (49).

Even today, PNP testing is still considered as the norm. However, there is a strong interest and need to move and explore the use of CBT and multimedia in assessment

to adequately reflect what is delivered in the instructional design of teaching. It is time to think of testing as a learning tool in addition to it being an evaluation tool (80).

Just recently in 2017, two articles were published regarding video use, one article by Dong and Goh (2014) outlined twelve tips on how to reduce video production time and increase their reusability. These tips were on the effective use of videos in medical education that were based on a review of research done on the use of videos in education. They found that videos stimulated curiosity and because it situated the learner in a realistic clinical scenario, it engaged their attention and promoted authentic learning (132). The other article was examining the general trends and results from published articles that examined video usage in medical education (133). To have a general look at the history of multimedia Figure 3.1 presents a timeline that outlines the history of multimedia from the early 1960s.

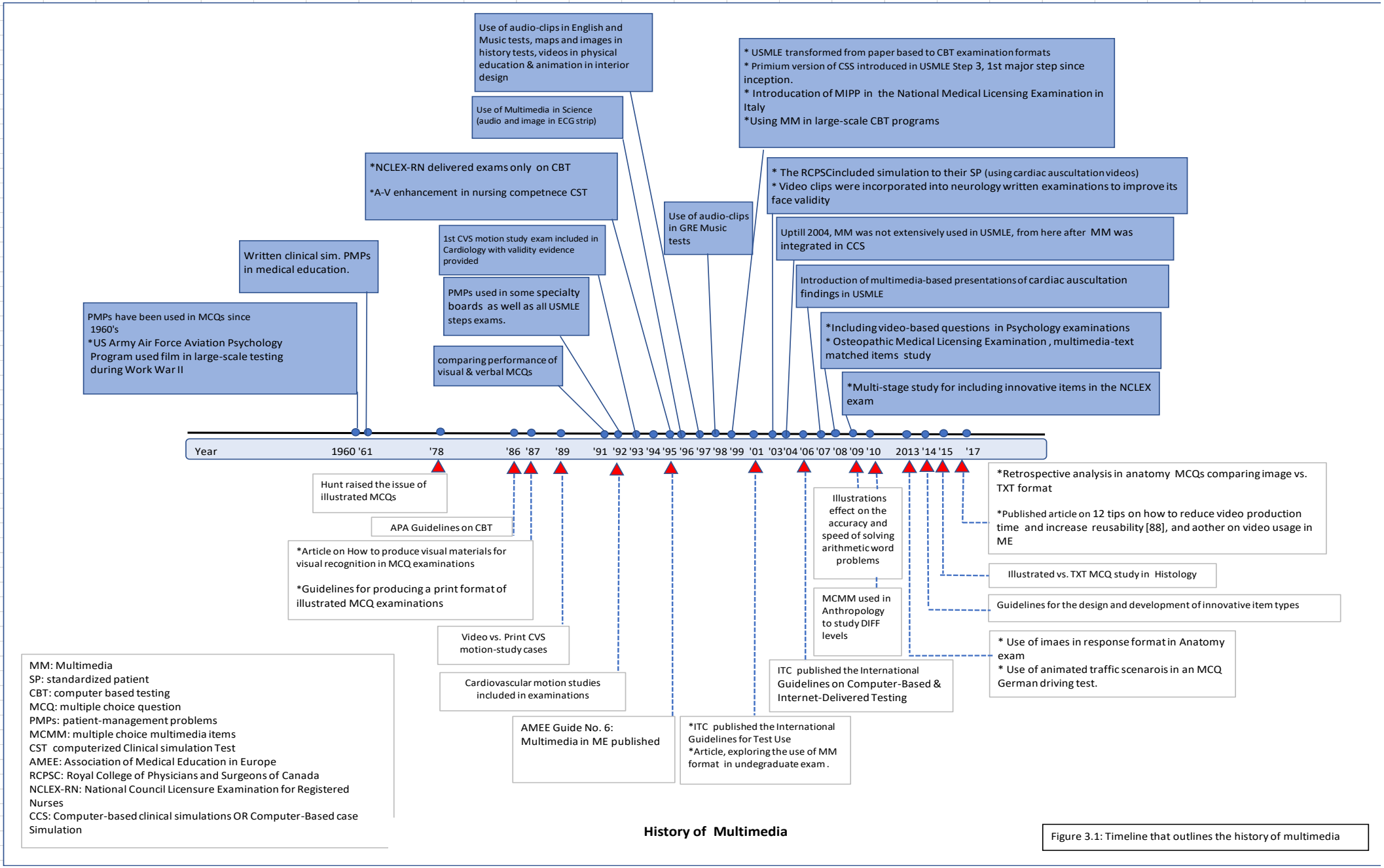


Figure 3.1: Timeline that outlines the history of multimedia

3.3.5 Issues associated with bringing multimedia to operational tests

Multimedia can measure important elements that conventional methods alone cannot assess (5, 7) and when they are introduced as a medium in examinations their purpose must be justified and several issues should be explored to ensure validity, alignment with the framework, and purpose of the study. When constructing an examination, the question that must be asked is: what are we trying to measure when introducing an image or video into an item within a given domain? This helps to shape which multimedia is appropriate for selection if at all needed. The presence of some multimedia might unintentionally distract the examinee and, hence, distort the intended construct that is being measured. This may occur if the used multimedia in a given item fails to capture the core component of a case or situation, under-represents it, or introduces construct irrelevant factors (7).

The use of multimedia may enhance an item and, therefore, measure an important aspect that wasn't possible in the conventional paper-and-pencil medium (5, 72). This was explained earlier by Bersky (1994), where nursing examinees who were subjected to the AV-format items felt more involved with the patient's in the cases and committed fewer risk and inappropriate actions than those who took the paper-and-pencil format (72). However, in another study by Shea et al. (1992) both the multimedia and conventional format were found to appear equivalent (19). In such cases, when both formats test the same construct, the justification for the use of multimedia should be to allow for a broader domain coverage, demonstrating a more appropriate alignment with what is happening in the clinical setting and what is conveyed in the teaching sessions (7).

3.3.6 Multimedia classification and issues related to the testing process

It is apparent from the literature that each research study is trying to understand the complexity of the dynamics that multimedia brings into learning and assessment with the various proposed methods of classifications and structures. Parshall, Stewart, and Ritter (1996) described how test measurement can be improved through the use of innovative test items through its ability to test something different. Their research described a taxonomy composed of five-branches for item innovation, which was later classified into seven dimensions for technology-enhanced items (TEI) (95). These TEI can be ordered on a continuum extending from least innovative to most innovative, and the more innovative an item is, the more it is dependent on complex computer functionality (99). Ruiz, Cook, and Levinson (2009) classified animation along various dimensions into nine categories under four domains. The four domains are a) nature of the process that is visualised; b) the learner's level of interactivity; c) dimensionality and; d) level of abstraction (38). Gulikers, Bastiaens, and Kirschner (2004) presented a theoretical framework that consisted of five dimensions for authentic assessment, which can vary in their degree of authenticity (117). The purpose was to provide guidelines for authentic implementation in computer-based assessment. A brief overview of these proposed frameworks and dimensions are explained in Appendix 3. According to Bennett et al. (1999), issues associated with including multimedia into operational tests can fall under four categories: Measurement, Test Development, Delivery, and Financial Categories (7). All headings have further subheadings and explanations on how these affect the use of multimedia. Liu, Papathanasiou, and Hao (2001) also proposed factors in the form of questions for a successful implementation of multimedia testing that was categorised into four headings: Students' Perspective,

Faculty's Perspective, Content and Technical Issues (80). Both studies proposed a series of overlapping questions under each heading that needed to be answered in order to develop a proper multimedia assessment. The former had proposed his view from a practical sequential stance while the latter focused on the stakeholders' perspectives. Appendix 4 outlines both views combined for a more general view that seems to make more sense. Further research in these areas is still needed.

The quality of multimedia materials is more than just being clear, it also involves a number of additional factors that were still not mentioned and were related to how multimedia was selected and written in the first place, as well as other factors affecting the nature of multimedia materials and how the materials were perceived by the learner. For example, the presence or absence of colour in an image may help identify landmarks, discriminate information, and contribute to the realism of an image (123). The size of an image presented will aid in identifying and clarifying what needs to be interpreted to answer the question. When selecting multimedia, attention should be directed to the size of materials selected particularly when using the enlargement function (zoom in) to avoid that potential loss of details in the image and the inability to recognize important anatomical structures that are present but are not clear (7, 56). In addition, the degree of difficulty and complexity of the context of an item may direct the examinees' attention to the image to search for more information that is not readily understood from the text (123). The computer and software interface play a role in interpreting the information from the item, as examinees are given the chance to experience with a free navigation system that uses branching or embedded links to allow users to move or jump from one part of the examination to another (for e.g., mark unanswered items) (94). It also allows them to practice other functions such as to

enlarge, rotate, zoom in or out of an image, replay a video, adjust and control for audio clips as needed at their own pace and readiness (80, 94, 95, 106, 134).

Another issue that should be taken carefully, which is related to the individual's characteristics, is the issue of fairness. Examinees with hearing impairment or visual disabilities should be considered when preparing any multimedia materials. To ensure fairness, examinees with disabilities should be of particular concern when incorporating a video or audio in a question. This could be solved in some cases by the inclusion of closed captioning or using videos that do not contain audio at all (7). For those with visual impairment, they might need an audio-only case or a narrator that describes what is happening, although this might not always be feasible (7). Motivational issues may also play a role when using multimedia. For some groups, this may have a positive effect, as they may view exams with audio-visual materials to be more interesting and familiar(7). Other factors that contribute are the age of the individuals, level of experience in training that affect ability, language and reading ability, visual literacy, and the skill and willingness to read visual materials (123). All these may interfere with identifying relevant information from irrelevant ones and cause the individuals to shift their attention from one format to another (123).

One of the major issues in the test development process that make incorporating multimedia difficult is obtaining the source and materials for it. Creating one's own original material is cumbersome, time-consuming, expensive, and requires different speciality backgrounds with different talents to review them (7). Another matter is the copyright issue and obtaining permission for use and publication of such materials, which could be quite tiring (7). One solution would be to have a general agreement or contract with a large multimedia library where segments and extracts from library entries could be made (7), but even this solution could prove to be difficult for many.

In the process of delivery, concerns are raised about handling and transferring multimedia electronically. The issue of size, as well as security, are of major concerns. Regarding size, multimedia files could be fairly large, having a two-minute video clip may exceed in size an entire testing bank (7). This leads to the issue of security, as transfer would be conducted through a compressed file in a USB drive or uploaded to an online database, such as clouds and online drives, which could alter some of the features in the materials, risking the spread of items, as well as viruses. Another important delivery issue is the technical and logistic issues related to multimedia. This includes upgrading test centres with the necessary capabilities, having appropriate and up-to-date software and hardware that are regularly checked and tested by test administrators and IT specialists to ensure the smoothness of multimedia displays and functions during an examination (7). This sometimes requires the availability of staff during after-hours working time and during the weekends when examinations are sometimes scheduled for a quick onsite response. Other materials that would be needed are headphones, sound mufflers and examination logistics (registration area, computers, stationery, etc.). One major and crucial aspect of test delivery is the availability of a large storage capacity unit that is not connected to internal networks or the internet, and is able to store, backup, and handle the large electronic test files and materials. All these and more would add up to the cost associated with incorporating multimedia in computer-based examinations (7).

This leads to financial issues related to multimedia and test development. The cost of such a high-stake large-scale examination is burdensome, particularly if conducted on a frequent basis (7, 97). Additional costs include reserving seats for registration, source of materials, item writers' test material, reviewer's time and effort for the review and quality check process, delivery, publication, transmission, and storage of items.

All these issues add on to make the decision to process with such an examination, and should only be followed through when delivering multimedia exams benefits and exceeds the expected costs (7). Although it is believed that these exams generally take longer to develop, they will have a longer shelf-life and will be easier to revise later on (80). In regards to multimedia use, the issue at hand is not only concerned with if multimedia would work or not but also under which circumstances would one need to use them and how would it be effectively used. Therefore, in the following section, we will take a look at how the brain processes verbal and visual information from instructional materials, and how this can help in understanding the characteristics of multimedia and, in turn, would aid us in applying these principles into assessment.

3.3.7 Cognitive load theory, multimedia learning and instruction

The literature is limited with research in the field of medical education regarding the role of multimedia in assessment or the use of instructional computer animations, and the ones that are available give contradictory evidence (38, 54, 119). Fortunately and for many years, the non-medical education literature has researched the area of design and multimedia use (38, 135). However, most of the research regarding cognitive load theory, multimedia (visual representation) and how verbal and visual materials are processed were in the area of teaching, learning, and instructional design (53, 54, 114, 119). This is because the use of multimedia and digital resources is considered a central component in the learning and teaching environment (38, 136). Furthermore, most of those researches have fixated on the effect of images on memory or the transfer of learning content from one format to another (54, 137). Yet still, with all these advances, research on how these images are used within written assessment lag behind (53, 119). This is important to understand in order to be able

to transfer the knowledge and use of these findings into assessment and be able to further understand its psychometric characteristics and interpretations (54, 114).

There are many theories that report about the use of verbal and visual displays in learning and offer guidelines for instructional design, the most dominant one is the cognitive load theory (114, 135). The cognitive load theory, as explained by Cook (2006), 'provides a theoretical foundation for designing instructional materials to best enhance learning.' (114). The basic premise of this theory is that learning will be hindered if the instructional materials overwhelm a learner's cognitive resources (114). Further explanation will be given on multimedia learning and the integrated model for the comprehension of text and picture.

3.3.7.1 Cognitive load theory

The cognitive load theory provides insight into how the brain processes visual information in working memory and offers insights on what difficulties a student may encounter when learning new materials. It suggests how visual materials should be designed in order to optimize their effectiveness and reduce unwanted cognitive overload (114). The idea of cognitive load theory resides in trying to minimise the effort required for learning (138). The cognitive load theory consists of three types of memory and three categories of cognitive load. The three *types of memory* form what is called the "cognitive architecture" and consists of sensory memory (SM), a limited working memory (WM) and an unlimited long-term memory (LTM) that stores knowledge and information permanently (54, 114, 124, 139). According to Paas, Tuovinen, Tabbers, and Van Gerven (2003) cited in Berends (126), the three categories of cognitive loads are: the germane load (that deals with adding useful information and building of cognitive schemas), the extraneous load (that deals with

processing irrelevant information), and the intrinsic load (that deals with the amount (volume) and level (complexity) of information to be processed) (126). One of the goals of instruction in cognitive load theory is to maximise germane load and minimise extraneous cognitive load, therefore, freeing working memory to enable learning and understanding (38).

3.3.7.2 The working memory (WM)

The working memory is made up of two components (or processing systems) that process visual and verbal information independently, a visual-spatial sketchpad and a phonological loop (114). This means that visual and textual materials are processed in different cognitive systems (120). It is assumed through information processing theories that people have limited working memory, and so learning will not take place when it is overloaded (114). The working memory has a limited capacity to hold and manipulate words and images of approximately six “units” of information (54, 124, 139) and (97, 135). What determines how much information is retained at the same time in working memory is the learner’s prior knowledge (114).

The working memory can also be affected by the nature and medium of the material presented. For example, instructions with visual representations can burden the working memory’s limited capacity (114). Therefore, the weight placed on working memory can be reduced by reducing its cognitive load or by increasing its capacity. This is important to note, as information might overwhelm one of these processing systems. If the amount of cognitive load exceeds the capacity of the working memory then performance will be less accurate and slower(126). Understanding this can help in managing the information being divided across these systems instead of

overloading one or the other (114). This can be used in increasing the capacity of working memory by using more than one presentation modality (114).

3.3.7.3 Multimedia learning and instruction

In order to have a holistic view of the impact of visual representation (multimedia use), we must first have an understanding of visual representation in science education and what factors affect it. Research has shown little in this area as most research has focused on verbal learning, However, it does show that instructional representations should be designed in a way to reduce cognitive load as learners have limited working memory (114). The following section explains the characteristics of words and pictures and what is meant by a mental model, followed by a brief description of multimedia learning before explaining the information processing system.

3.3.7.3.1 About words and pictures

Words (or text) consist of symbols that have a random structure and are only comprehended if their culture and rules are understood (54, 111). They are known as descriptive representations when described, for instance, in a scenario where one would use nouns, verbs and prepositions to refer to its parts and relate them to each other (111). Words can be presented in two forms either as a) printed on paper or text on-screen or b) as spoken (verbal narration) (139). Spoken texts (auditory text) are direct meaning and, therefore, in this case, students cannot jump back and forth as they would in written text (114).

Pictures (or images), on the other hand, are quite different from words and consist of iconic signs. They usually represent something real and are related to an object by similarity (54, 111). Pictures, whether being multimedia or visual displays, are considered depictive representations. When presented, information about the size,

shape, and form of an object is gained all at once (111). They can take different forms and can either be static presented as (photos, images, illustrations, charts, diagrams, graphs, or even maps) or be dynamic as in (videos, animations and interactive illustrations) (99, 139). Static representations (or visualizations) are precise single-framed that are taken from a flow of motion with no affected change in them with respect to time. While dynamic displays (or visualizations) represent a continuous flow and consist of a series of frames or (depictions) that are changing continuously over time (99, 140). Most multimedia research make their predictions about learning using text and visualization, but not on the different visualization formats (static and dynamic) (140). And those that compare static with dynamic visualization reveale inconsistencies in their results and most, if not all, recommend considering when and why to use a certain type of visualisation (140). To further elaborate how pictures and words differ in their presentation if a characteristic of a rash was to be described through text, then perhaps only specification of its form might be provided, eliminating any further information regarding size and orientation. However, when this is presented in a picture then the information received regarding form, shape, colour, and size are not limited (111).

Pictures are classified into five categories according to their functions as: decorational, representational, organisational, interpretational, and transformational (141, 142). Decorative pictures, as the name entails, are not related to the text and therefore, do not contribute to understanding it. Whereas representational pictures illustrate the context and situation of the text. Organisational pictures illustrate steps of tasks that are described in the text while interpretational pictures are simple illustrations that aid in explaining complex systems. Finally, transformational pictures augment the recall of text information (141, 142). Therefore, it is important to understand that pictures do

function differently in test items and their perception and interpretation may differ among individuals in relation to their associated text (142).

To take matters forward, there are two main visual information that affect the comprehension of multimedia in an item: the visual context and the visual content. Context visuals (also knowns as visual representation) provides contextual information (i.e., information about the context components and setting). Ginther, (2002) cited in Wu (143). These context visuals embedded in the test item may affect the reader's familiarity with the context, and would either increase the difficulty level of an item because it acts as a distracting or misleading representation, or it may decrease item difficulty and increase its validity (143). Further, research is needed to understand the use of multimedia and context in CBT and its relation to students' performance (143).

The other information that is important is the content visual of the task that is "related to the content of the verbal stimulus" (Ginther, 2002, p.134 cited in Wu (143)). Information from the content visual could complement the item's verbal description and aid in solving the problem. In order for test-takers to be able to answer a question, they would need to retrieve and comprehend the necessary visual context and content information from these visual representations (143).

Video and animation cannot be said to be superior to static or still images. Depending on the content and context of the situation, they may impede learning and assessment (38, 99). Researchers should address factors that may affect these formats such as exam content, individual characteristics (age, level, spatial ability), cognitive demands, and so on (99). Questions such as when to use static images versus two-dimensional media versus three-dimensional animation? And what are the educational impacts, costs and benefits of using multimedia? All need to be investigated.

3.3.7.3.2 Mental model

Multimedia (or visual resources) play an important role in developing a student's mental model, as students may view visual representations in different ways (120). Some students observe themselves as being visual learners and appreciate the use of visual resources in questions. Therefore, it is important to explain the concept of the mental model (also known as mental representation), which simply means a reader's own personal understanding of the text that is based on his/her prior knowledge and ideas already known to him/her (120). Most of this process is automatic and unconscious and is built as a result of processing the text when reading a question (120). The representation formed is a collection of concepts, images, emotions and the relationships between the concept and not the actual words (120). Therefore, each reader may develop their own personal understanding of the text and, hence, their own mental model to the text based on what seems to be relevant to them. Consequently, readers may not all have the same representation (120). Constructing a mental model from text alone requires some effort and may be subjected to misinterpretations and affect one's comprehension because the text would need to be interpreted, unlike pictures where a mental model can be constructed more directly from them. Therefore, presenting a picture with text helps readers construct an initial mental framework from the text (121).

The act of learning from words and pictures is called 'multimedia learning'. While presenting these words and pictures to produce learning is known as 'multimedia instruction' (139). One of the main challenges that designers of multimedia instruction face is something called 'cognitive overload'. It is defined by Mayer and Moreno (2003), as: 'the learner's intended cognitive processing exceeds the learner's available cognitive capacity' (139). In other words, it occurs when the cognitive processing

demand from a learning task exceeds the processing capacity of the learner's cognitive system (139).

3.3.7.3.3 The theory of multimedia learning

A derivative of the cognitive load theory is the cognitive theory of multimedia learning (138, 139, 144), which sets the principles for designing educational materials (i.e., multimedia presentations) that facilitate knowledge retention and deep meaningful learning for medical learners through the use of lecture slides enhanced multimedia (138). The cognitive theory of multimedia learning provides insights into how people learn from pictures and words (144). Because multimedia learning principles are based on cognitive load theory, the ideas are grounded in the work of cognitive psychologists Mayer and his colleagues who extensively studied and established their work in the cognitive load theory (138, 144) and described basic and advanced principles of multimedia learning and presentations. They explained the appropriate use of text, images (still and dynamic), and audio materials for instructional purposes (138, 144).

The theory of multimedia learning is based on three assumptions on how the mind works. The first assumption is called the (dual-channel assumption), which proposes that humans hold two separate systems (or channels) for processing verbal and pictorial material (or information) (139). These channels are subjected to a second assumption, which is called the (limited-capacity assumption). This assumption means that each of these channels has a limited capacity for the amount of material and information that they could hold and process at one given time (139). Furthermore, cognitive processing occurs by building a connection between verbal and pictorial representations in order for meaningful learning to occur. This is the third assumption and is known as the (active-processing assumption) (139).

3.3.7.3.3.1 The processing of information: How does it occur?

The human information processing system comprises of two separate channels named the dual-channel theory of multimedia learning that has a limited capacity for processing incoming information (139, 144). The first channel processes visual input and information presented in the pictorial or visual format while the second channel processes auditory input or information presented in the verbal or auditory format (139, 144). An extensive amount of cognitive processing takes place in these channels for meaningful learning to occur. (139).

When new information (from the environment) is presented to the learner in a multimedia presentation, that consists of both words and pictures, it has to go through active processing steps of selection, organisation and integration to make sense of the instructional material received (139, 140, 144). When knowledge is presented to the learner, it is transformed into five different modes throughout this whole process (139): 1) physical representations (words and pictures); 2) sensory representation (to the eyes and ears); 3) working memory representations (sounds or images); 4) deep working memory representations (verbal or pictorial models); and, finally, 5) long-term memory representations (stimulating relevant prior knowledge and building links between the verbal and illustrated information) of the learner (139, 145). This can only be achieved if both sources of information enter the working memory at the same time (111, 145).

Words, as previously explained, can be presented in the sensory memory in two forms, either spoken words that are heard or words that are read, and therefore, they can either be processed in the *verbal system* as descriptive (represented as spoken words impinging on the ears) or in the *imagery system* as depictive (represented as printed

words impinging on the eyes). Pictures, on the other hand, are processed in the imagery system (represented as pictures impinging on the eyes) (111, 139, 145).

When the learner pays attention to the auditory sensations coming from the ears he/she selects words in order to process them onwards from the sensory memory to the working memory, in the same way the learner pays attention to the visual sensation coming from the eyes and selects images to process them to the working memory (139). When in the working memory, the selected words can either be processed as verbal information (sound) if coming from the ear or as pictorial information (e.g., an image) if perceived by the eyes, while the selected pictures are processed as pictorial information (139, 140, 144). These sounds and images are attended to in the working memory and are organised by the learner through constructing a verbal mental model from the incoming words or a pictorial mental model from the incoming images, respectively (139, 140, 144). As long as the information is attended to in the working memory, it stays there (54).

Finally, for a deeper understanding of the content by the learner, a connection (encoding process) between these two mental models is built. At this stage, relevant prior knowledge stored in the long-term memory (LTM) (144) is integrated with the two models to elaborate their inter-representational connections and form an integrated mental model (139, 140, 144) and information is stored in the LTM. When one wants to remember something, the information from the LTM is retrieved into the WM again and remains there as long as it is attended to (54). The integrated mental model is the end product of how an individual understands text and pictures that are presented in different forms (145).

This was a brief explanation of the cognitive load theory and dual-channel theory for multimedia learning that needed elaboration in order to be able to understand what factors can play a role and affect multimedia in learning and, hence, would assist us to understand its functions in testing.

There are two other aspects that one has to consider when trying to comprehend the role fo visual representation: 1) the way that they are designed and 2) the way that they are interpreted by different learners (114). The effect on cognition doesn't rest alone on how one interacts with visual representations alone; other factors such as an individual's prior knowledge (as part of individual differences) spatial ability and other factors are important when trying to understand the impact of these representations on a subject's cognitive structure and process (38, 114). The following section further explains these in detail.

3.3.7.3.4 Multimedia principles of instructional design, theories and factors affecting the cognitive load

It is apparent now that in the cognitive load theory, the limiting factor in cognitive tasks is the working memory capacity (54). Also, in multimedia theory, the main risk of adding multimedia to text in the context of a test is that it would lead the reader to form a mental model (representation) of the test question that might not match the meaning that was intended by the item writer (120). Effects of adding multimedia to text-based testing materials were discussed in the literature and demonstrated that pictures affected the cognitive process in testing materials in a similar manner as they did in learning (121, 146).

Extensive research to understand how to use words and pictures to foster meaningful learning has been undertaken by Mayer and Moreno (2003). They defined meaningful learning as 'deep understanding of the material, which includes attending to important

aspects of the presented material, mentally organising it into a coherent cognitive structure, and integrating it with relevant existing knowledge' (139). It is reflected by applying what was learnt into new situations and can be measured using retention tests (recalling what was presented in a lesson) or through transfer tests (solving problems from the presented materials) (139).

Mayer's principles for designing effective instructional multimedia materials encourage educators to design the instructional message in a way to enable the learner's cognitive learning process. It contains many sub principles that the educator should incorporate and are centred around three ideas: 1) eliminating external distractors; 2) encouraging learners to establish mental frames from the presented materials' and 3) facilitating the integration of prior knowledge with the new material (144). For example, educators should eliminate external distracters through the removal of extraneous words, pictures, and sounds, and facilitate the learner to integrate new knowledge and materials with prior knowledge from other disciplines by establishing what the learners already know (144). Educators should encourage learners to establish a mental frame for the materials by making objectives clear to them beforehand and encouraging them to prepare for the topic beforehand, this will aid in reducing the cognitive load and not overloading the learners' cognitive capacity with information (144). The main principle that concerns this research is Mayer's multimedia principle, which recommends presenting both words and pictures together rather than presenting words alone when describing material (144). The literature also covers other multimedia factors, principles, and theories that may cause cognitive overload and altering the information, as well as those that may reduce the cognitive load. These are further explained below as follows:

3.3.7.3.4.1 Paivio's (1975) dual-mode presentations

The dual coding theory informs us that visual and verbal information are processed independently in systems of working memory each creating a mental model and later are mapped onto each other, and that more information is processed and maximized when using both system capacities (text with graphic) rather than either of them separately (114, 120). This agrees with Mayer's multimedia principle discussed above (144). However, this should be used when visual and verbal information is incomprehensible in isolation, removing the appearance of any redundant material (see section 3.3.7.3.4.7). Visual resources have a large role in developing a student's mental model and are considered superior over text because their general meaning can be quickly grasped, they require less cognitive effort, their elements are processed simultaneously rather than sequentially as in text, and they are double-coded when processed, once as an image and another time as verbal labels, whilst words are only encoded verbally (120). This may lead to bias towards information that is gained from visual resources. In other words, it is richer to learn from a text with graphics than with either of them alone. It must be noted that pictures can sometimes have a negative effect on the mental model construction (114), and some view that presenting both image and text may be harmful as they cause a split in attention between the two forms of information that then have to be integrated (120).

In summary, when the task mostly requires the use of either the pictorial mental model or the integrated mental model, then one would benefit from using the integrated text and visualisations over text alone resulting in higher learning outcomes (38, 112, 140, 144) Hence, the multimedia principle of instructional design that is explained by the cognitive theory of multimedia learning (140, 144).

3.3.7.3.4.2 Prior knowledge

It is significant to note that the role of individual differences, particularly prior knowledge, is an important determinant of learning and in understanding how visual representations and their design impacts learners' cognitive structures and abilities (114). Learners construct their concepts from prior knowledge; they use this knowledge to select relevant information from multimedia and add information from their prior knowledge to eventually develop a mental model. Prior knowledge also influences attention and perception. All this contributes to the variations on how learners interpret visual representations (114, 123). The learner's level of knowledge and expertise in a content area may affect their performance with multimedia (38); the less prior knowledge an individual has, the more likely they are to be subjected to cognitive overload (114). Prior knowledge can determine how easy a learner can interpret and perceive visual representation in the working memory (114).

3.3.7.3.4.3. Spatial ability and orientation

Another important aspect in understanding multimedia is the concept of spatial ability, which is the ability to be able to understand the position, structure, and manipulation of an object in a three-dimensional structure (38, 54, 147). Spatial ability differs from visual learning style and is defined by Peters et al., 1995, cited in Vorstenbosch (50) 'as the ability to rotate images mentally' or as defined by Kozhevnikov, Hgarty and Mayer 'the ability to mentally interpret the spatial relationships between parts of an object or between different objects in space' (148). For example, imagining the heart in its correct position. Visual information about the characteristics of an object and spatial configurations are handled in different cognitive subsystems. One system is used to identify the object through knowledge about its appearance, and the other

system tries to locate the object through knowledge regarding distances between objects and spatial directions (111). Depending on the multimedia type, as explained, certain visualisations may reduce the processing demand, as examinees do not have to mentally imagine and animate the spatial changes of a situation or procedure on their own (140). Moreover, there is a risk of incorrectly reconstructing the changes mentally. This is true regarding the use of dynamic visualisation in the learning process. Spatial ability may be a factor in the way visual information is processed and how easy it seems to the reader (50). This concept is important because the viewpoints from which an object is seen may differ, as the learning process can be affected by the study of two-dimensional materials (illustrated) or three-dimensional material (real) (147). In addition, students with higher spatial ability are more prone to get higher scores in examinations (54). However, weak students who are unable to have the ability to animate and spatially orient the situation mentally seem to benefit more than higher ability students from the multimedia items than text description items (140). It should also be noted that low-ability students have often been found to have deficient working memory capacity because they reach their maximum load sooner than high performers or experts would (126).

Another important aspect that relates to the content of the item is the cut of an image and its relation to other structures. Some studies show that transforming information from a cross-sectional view to a normal view requires extra effort in working memory (54). Most images when displayed have a fairly uniform pattern except cross-sectional images. It is important to recognize the subject of spatial ability and the information it yields because it has an implication on the trainee’s selection and performance (147). Also, when considering that spatial ability may take effect in several items, then it is

assumed that these small effects on these individual items compiled will have a significant effect on the total testing score (54).

3.3.7.3.4.4. Position of multimedia in relation to text

From a cognitive load perspective, integrating information from text and pictures can overload the working memory, leaving fewer resources to solve the original task (126). Therefore, to reduce cognitive load that requires integrating multiple sources of information, multimedia and text should be presented simultaneously (placed at the same time), and printed words should be presented next to the corresponding multimedia (111, 114, 144). Again, this reflects two of Mayer's principles, spatial contiguity, which states that printed words should be placed next to corresponding graphs, and temporal contiguity, which states that the corresponding narration and animation should be placed at the same time (144). When this is not possible, it is recommended to present the picture first followed by the text (111, 123). Presenting pictures prior or with the corresponding text facilitates recall and comprehension (121), activates prior knowledge, and provides a scheme for organising the following text material (123). While presenting the picture after a text may not yield any beneficial effect or may even have an unfavourable effect on comprehension. This is called the 'picture-text sequencing effect' (Schnotz, 2014) as cited in Linder (121); where the picture may interfere with the mental model constructed from the text, diverging it from the pictorial mental model causing confusion during item solving and affecting test performance.(121). In addition, working memory requires little space when processing pictures first, leaving enough space for processing text. However, if texts were processed first, then most of the working memory capacity will be used and little space would be left to process the picture (111). Processing the text first would lead to the developent of a mental model based on the text format, which might be different

from the presented picture that follows and, therefore, would interfere with the picutre when being processed and interpreted (111).

3.3.7.3.4.5. Animated multimedia (animation)

Animation is a way of presenting a dynamic phenomenon that is perceived in the physical world by using multiple images over time. Results from research using animations are inconsistent, with some providing its benefits over static graphics and others not. Animation differs from static images in that when using a static image, the student needs to extract the relevant information from the image and include it in a schema and construct a mental model around it. However, in animation, this is more complex because of the dynamic nature that makes it difficult for the student to extract relevant information from a fast-paced stream of animation, which may require successive viewings by the students in order to be able to have enough time to interpret the animation. This can all affect and overload the cognitive load (114). To elaborate further, both verbal and pictorial channels are overloaded in the working memory in the following situation: if the content of the information is rich (in other words complex) and the pace (rate) of the presentation (e.g., video) is fast, then the learner will not have enough time to get involved in the deeper cognitive processes (i.e., organising words into a verbal model, and images into visual one, and later integrating them) (139). This indicates that sometimes a simple graphic that allows the student to extract the relevant information from it might be a better use than a realistic animation. One can overcome this problem in animation by having interactive controls where students are able to review segments of the video (114). Video clips that are chosen should be selected, and developed in a way that the student grasps the whole rather than the details of the clips. As clips that are focused on details will require multiple

viewing for careful analysis, which might be distracting (122) and would need to be factored in as additional time allowed for the examination.

It is also important that information load from multimedia should match the time selected for the educational and testing activity (140, 144), and when using animations it should be paced according to the learner-paced segment (144). This is in accordance with one of Mayer's principles "segmenting principle" where an animation should be offered in learner-paced segments (144). Although it might not be possible to simplify a presented video, it is possible to slow down the segments by keeping information on the screen for longer (e.g., by increase looping time). This is called tracing (38, 139). Another way is allowing for more time between successive segments (loops) and, in some cases, it might be required to segment the materials to present them in chunks (139). In a research study that used videos, participants did not stop the videos that contained looping (video playing over and over again) (39) and most mentioned that the video-clip loops distracted them and hampered their performance (39). Moreover, it is recommended that video clips be looped only once, or students are given the option of choosing whether to loop or not (39). From the reviewed literature, there were no specified comments or recommendations found regarding how long a video-clip should be except for one study conducted on behavioural data, which recommended that video clips should be 2-3 minutes in length (122), and another study that suggested from their results that multimedia items required, on average, 30 to 60 seconds longer for a response time than text versions (20). Test developers should remember that if a big proportion of the exam items consist of multimedia, then the examination time should be increased or adjusted and further be studied well (55).

3.3.7.3.4.6. Frame selection

It is important to remember that not all video materials can be shown in a print format because it is not possible to capture dynamic motion in still frames. Attempting to do so may lead to additional confusion or misunderstanding from the examinee's side (19). Additionally, within the print format and for some cases, selecting a small number of frames may act as a clue and direct the candidates to the correct answer. Therefore, for each case, in an attempt to avoid cuing, a small selected number of frames should include both significant and non-significant views (19). In some cases, selecting certain frames from a motion study imposes a slight limitation, as it cannot convey all of the important clinical information that is needed for detecting or identifying some abnormalities. This may lead test developers to exclude some disorders from the print format when selecting items for examinations (19, 130). However, if selected still-frame images of the cases could be used successfully in the print format, it would reduce the cost of developing and administering the AV format. When using more than one multimedia format in a single question, it should be taken into account that examinees must use this information to correctly diagnose the patient's case and that the correct diagnosis might hinge on one of the multimedia information (for example an audio-clip) (7). It is important to also note that videos and animation are not intrinsic properties of still images, and therefore, corresponding and additional information need to be extracted from the accompanying text and aligned with the multimedia. This process might be viewed as resource-intensive and prone to error (140).

3.3.7.3.4.7. Irrelevant (redundant) multimedia

The content of the multimedia item could either add or remove information relevant to the item scenario and, hence, make the item either easier or difficult for the examinee.

Multimedia content should illustrate key instructional points and minimise irrelevant information to reduce a student's cognitive load (124, 135, 139). The MM content also could change the examinee's response pattern in the same topic when given in a text or multimedia item (5). It could also have the effect of accidentally distorting the intended construct under measurement. This occurs when the multimedia stimulus does not capture the essence of the construct or introduces irrelevant factors (5) (e.g., an ultrasound is not clear because it is too zoomed in). Nevertheless, multimedia could equally capture the essential elements of the construct that is not captured in the conventional way (5). If any type of multimedia (e.g., an image) is not needed in the question, the examiner is required to balance the advantage of including the image to become less daunting to the examinee against including it to distract them away from the text and intended purpose of the question (120). This coincides with Mayer's principle called the coherence principle that calls for excluding extraneous words, pictures, and sounds to eliminate external distractors (144).

Information can be maximizing when irrelevant information is avoided (38). Including irrelevant multimedia next to a written clinical scenario increases the extraneous cognitive load which might result in longer response time (126). The inclusion of visual resources that are not deemed necessary was expressed by some students in a study as being reassuring. However, it is still recommended to exclude this extraneous information when designing the materials and highlighting the important aspects, as Mayer's signalling principle proposes to highlight the essential materials (120, 144). Moreover, if the same information is presented in two different methods (text and image) then the student is required to process the information twice. This means that if an image can be presented and understood by itself, then adding explanatory verbal information (e.g., to restate what was in the image) would be redundant. Duplication

or overlapping of information in text and graphs leads to the unnecessary burden on the working memory and cognitive resources (114, 126) and has a detrimental effect on performance (49) because learners' attention is split between two visual sources (the text and picture) (114). This also can be applied when writing items, where the item writer should balance for the need to include duplicated information or not (120, 123).

3.3.7.3.4.8 Cognitive schemas

Another way of reducing the cognitive load is through the construction of cognitive schemas (114). Schemas that are stored in long-term memory help organise a large amount of information processed as a single unit and can link relevant information together that is accessible when needed (114). A number of instructional design guidelines (114) have been proposed to facilitate schema construction by reducing working memory load and some are mentioned below:

3.3.7.3.4.8.1. Multiple representations

Novice learners have difficulty coordinating and linking visual representations with their cognitive interpretation of the graphic. They spend much of their cognitive resource on interpretation and less resource for linking. In multiple representations novice learners switch between representations and do not make use of them. If this switching occurs, it indicates that the learner has difficulty understanding what is presented. Translating representations is difficult for them because it requires having an underlying knowledge of the concept (114).

3.3.7.3.4.8.2. Modality effect

The modality effect or modality principle as Mayer explained, is presenting words as narrations instead of printed text (144). Sometimes, presenting verbal information through narration instead of written text eliminates the competition for visual attention (114). As previously explained, both text and pictures are processed in the same visual subsystem of the working memory. This is because text initially competes with pictures when it is being processed in the working memory and may affect the reader to attend to relevant information (114). In this case, presenting verbal auditory information instead as text is useful, as it reduces unnecessary cognitive load because it is being processed in the verbal subsystem and, hence, is not competing with the picture (114).

3.3.7.3.4.8.3. Materials with interacting elements

If the materials provided have a high interactive level particularly for the novice learner, there will be an overload on the cognitive resources and learning will not occur. Therefore, to reduce the working memory load for the novice, highly interactive elements need to be presented in isolation (114).

3.3.7.3.4.8.4 Instructional guidance

Instructional guidance, as well as providing an explanation, is an important factor for visual representation as it provides assistance to the learners with little prior knowledge to complete the task required, avoiding unnecessary cognitive load to construct schemas (114). This is also reflected in Mayer's principle of pre-training, which means to prepare or read ahead of time (144).

Other factors for cognitive schemas that should be considered are learners' spatial ability and cognitive ability, developmental level, sociological perspective, expertise,

and, most importantly prior knowledge (114). To give an example regarding expertise, students with little prior knowledge (hence, little schema) have difficulty differentiating between information that is relevant and irrelevant, and therefore, focus on surface features when interpreting images, making it difficult for them to reach to the underlying principle of concept and developing a comprehensive mental model (114). Experts, on the other hand, are able to see beyond the superficial features due to their extensive prior knowledge and developed schema that makes them able to make sense of the underlying principle and develop a comprehensive mental model (114).

3.3.7.3.4.9 The Interface design principle

Computer and software interface play a role in how the material will appear on the screen and affects how examinees receive the information. Multimedia characteristics are not superficial and independent from the item or the examinee; in fact, it is the combination of all these that help produce the final result of the examinee. For example, when using limited screen space it is important not to present a lot of information on the screen to avoid overloading the learner, this is according to the interface design principle (80). Presenting one question at a time is considered appropriate, particularly if using multimedia materials with it; as it is mostly perceived to be less overwhelming, less stressful and makes testing more relaxing (39, 80). However, time may be lost if examinees try to scan the exam to get a feel of the items and the whole exam, as opposed to easily scanning the questions in the paper-and-pencil examination. (80).

3.3.7.3.4.10 Split attention effect

A learner can attend to reading materials on computer-based instructions and to the computer screen at the same time (39, 126, 139, 149). This means that they can split

their attention between decoding language and decoding pictures (39). This phenomenon, based on the cognitive load theory, is labelled as split-attention effect and has been reported to cause a cognitive overload on working memory and subsequently hindering learning (39, 126, 139). All materials should be presented in a way that facilitates learning. Information needs to be integrated from verbal and visual materials. If the image does not fit and coordinate with the verbal material, the integration will be difficult. Schema can only occur after integration takes place. And integration requires that all material fit and coordinate with each other so that the learner's attention does not feel split between two modes of information as also explained in the modality principle (114).

3.3.7.3.4.11 Level of reader (novice or expert)

Van der Gijp et.al. (2017) reviewed the literature of the past 20 years for visual search patterns by use of eye-tracking in the radiology domain to identify the relationship between visual search and expertise level and had found six emerging themes (150). These are as stated in their article: 'time on task, eye movement characteristics of experts, differences in visual attention, visual search patterns, search patterns in cross-sectional stack imaging, and the effect of teaching visual search strategies' (150). Their results were consistent with other studies and found that experts were characterized with a global-focal search pattern that helped them identify suspicious areas through a global search that is followed by a detailed focal search (150). This is because experts have a rich knowledge base, need shorter viewing times, and are able to know where to best look for abnormalities through their rich knowledge base foundation (150). Experts also have more domain knowledge (i.e., more schemas), which makes them more able to understand underlying concepts and principles when they are presented with visual representations (114). They are able to concentrate on

relevant information to construct an effective mental model, and when presented with novel information they are able to interpret it by using their relevant prior knowledge as a starting point (114). When comparing experts to novice learners, novice learners tend to have pieces of information that are weakly connected, also known as “fragmented knowledge” (114) and when presented with visual representation they tend to be focused on surface features because they lack a clear and integrated existing knowledge. Novices are also unable to coordinate features that are within or across multiple representations, which does not make it easy for them to develop an understanding of the underlying concepts (114).

3.3.7.3.4.12 The Effect of language format on the difficulty level of multimedia

Reading is required for completing any written examination. When constructing an MCQ exam, item writers should select which format is more suitable (text or multimedia), as well as write items as concisely as possible, limit sentence length, and control vocabulary level in order to limit the reading involved and reduce the cognitive processing demands on the examinees (151). For example, in a study that investigated the effects of multiple-choice listening tests on performance and perceptions between oral and written format, students felt intimidated and scared from long sentences and unknown vocabulary (151). With multimedia materials, the language includes not only words but also what and how to describe the content, what and when to depict and how these may affect the examinees’ perception of the item content and further on test performances. A research study showed that language seemed to change for examinees according to the format of the scenario in the presented item. The study showed that when the standard textbook terminology was replaced with multimedia material (sound or image) the multimedia item became more difficult for the examinee. This could be explained by the fact that the multimedia items

required a higher cognitive step (at least one more step) than the text-matched item in answering the question. This extra step was due to connecting the multimedia aspect of the item (for e.g., the sound of a grade-3 holosystolic murmur) to the correct terminology in the textbook. If the examinee failed to recognize the audio sound, he/she would not be able to answer the question even if they knew the implication of the murmur (5).

If the content of the item could be sufficiently described by text, or could be easily labelled by textbook terminology, then (keeping in mind the objective and level of the exam), the use of multimedia might not be favoured as it could cause the item to become more difficult and less discriminating (5, 120). A possible explanation for this could be found in the richness of the text with details, which made the item fair in the text format. In other words, all the relevant information needed that was depicted in an image or video could have been inferred to from the text format itself, which was sufficient enough (140). However, if the content of the item is not direct and is difficult to be adequately described through words, then the use of multimedia is advised to make the item easier and more discriminating (5).

It is important to review the selection of words carefully when writing an item, as some terminologies might be more familiar to students who study from textbooks than their synonymous words that are expressed by the item writer. For example, in an item that was given in the nursing examination, students were unable to understand the description of the area where an injection should be given ‘‘the inner surface of the forearm” while it seemed to be clear to the item writer who developed it (45). When reviewing the terminology in the textbook, as mentioned in the article ‘the term ventral (or dorsal) aspect of the forearm may have been the more correct and familiar term’

(45), while in the innovative format, it was more direct by presenting a picture with the potential injection sites.

When the content of an item involves motion or movement, then seeing and experiencing it through the use of multimedia makes the item easier for the examinee to answer than if it were presented in a descriptive narration of the movement, which is confusing and is perceived to be less direct (5). For example, demonstrating a hip flexion through the use of a video is more direct than describing it as “A 70-degree hip flexion with knee extension” to the examinee. In another study conducted on senior-level nursing students, students found that the text description of breath sounds to be unfamiliar to them and had to process it first before being able to complete the problem. While in the multimedia format it was more direct where students just selected the breath sounds (45). The nature of the multimedia simply gives out more information and more cues that assist in selecting the correct answer.This explanation is suggested in Vorstenbosch and Klaassen’s study (54). In their study, labelled images and answer lists were used in the response format in the speciality of anatomy. The use of an answer list contained only structure, to it when compared to the use of images that in addition to structure contains spatial relations and orientation (54) possibly delivering more cues to the examinee. However, having the list of anatomical names also gives out cues that are not present in images (54).

In addition, visual stems may give clues to key answers, thereby diminishing the interpretation component of the item. Particular care should be taken to avoid such clues, ensuring that the stem of the question is concise enough to direct the examinees’ attention to the structure of the question being asked (71). The use of multimedia may hint to the correct answer as certain terms or phrases used in an item may act as an unintended cue or code word to the underlying disorder (54). For

example, the use of the term “thumb-sign” may direct examinees to associate it to the diagnosis of epiglottitis. For a better understanding of this, further research is still required regarding the cognitive processes underlying these effects.

From the literature, we can conclude that multimedia-enhanced questions and computerized simulations (both low and high fidelity) are among the most promising practices for assessing higher cognitive skills and problem-solving skills in the medical field. However, in order to properly value the quality of the format, it would be useful to know how much time and effort was put in the development of the items and multimedia, and what form of validation was required to ensure quality (138). It is also apparent that further research is needed for this type of format and further refinements are needed regarding its application and validity implications (107).

3.4 Conclusion

After explaining this brief but complex section regarding cognitive load theory and multimedia learning theory, the question still remains, does multimedia influence assessment, and in what direction? We understand that in learning, multimedia has an impact on learning, and in assessment, it has an impact on the psychometric characteristics of test items. It is apparent that there is not a straightforward question and the answer lies in how examinees answer an item after reading the question (stimulus) and options (response format), how they take in the content, their readability level and much more, all of which have an effect on the complex retrieval task of answering a test item (54). This can be reflected through measuring item difficulty and discrimination (54) and understanding from the examinees how they perceived the items. Despite the available research on how we process both verbal and visual materials and despite the advances in cognitive theory, the evidence is limited on how

the inclusion of images and videos within written assessments should be and how to properly use them. Further researches and investigations are required (40).

The following chapter covers the concept of validity in assessment and validity frameworks that are used in the testing process and that are needed for the development of multimedia items in high-stakes examinations to ensure their quality and validity implications.

Chapter 4: Validity and Validity Framework

4.1 Introduction

Assessment in medical education is a broad topic and has been put up for discussion many times particularly among governing and accrediting bodies that are seeking evidence for clinical competency and proficiency (37). Governing bodies all around the globe want evidence of competency and proficiency. The public pressures such organisations to produce competent physicians (1, 47) and society expects educators to be able to deliver the appropriate tools in order to measure their physician's skills (107). Thus, there is much discussion about validity and assessment in medical education. Validity has been widely acknowledged in the test development process as being the most fundamental aspect. Without it, one cannot assert their interpretations of test results.

Assessment, as put by Andreatta and Gruppen (2009) '*provides evidence that a learner has acquired knowledge and skills within a field of instruction*'(37) p1029. As one gains more skills and knowledge, more rigorous, complex and multiple types of assessment methods and validity evidence are required for performance measurements (37). Different stages of learning and training require different levels of validity evidence. Formative assessments and feedback are sufficient enough at the beginning when learners start acquiring their knowledge and building up their skills, at this level content and process validity evidence are adequate. As learners go through their curriculum training and have gained experience and mastery of the construct through practice, validity evidence should further include internal structure and construct validity because it is used to make a decision on the learner's performance (37). At their final stages of learning, where learners must demonstrate their mastery skills in order to be able to make a judgment and decision on their

performance on the construct (high stakes), evidence of consequential validity is also needed (37). It should be understood that any assessment process should be an iterative one that is conducted carefully following a well-defined framework (152) and that not all forms of validity evidence are required to be gathered for one assessment procedure (37).

4.2 The Construct

Any validity discussion will have the abstract word “construct” evident in it. Assessment aims to measure some underlying construct. A construct is something abstract that can be described and is certain that it exists but is difficult to adequately define and, therefore, cannot be directly measured (37). Cronbach and Meehl (1995) defined a construct as ‘some postulated attribute of people, assumed to be reflected in test performance. In test validation, the attribute about which we make statements in interpreting a test is a construct’ (153). Andreatta’s definition is a simple one to comprehend and states the construct as ‘Something we believe exists and which can be described, but *which may not be amenable to direct measurement*’. Other words that may be used are hypothetical, imaginary, abstract, object, thing, or unit.

Examples of broad constructs in medicine are professionalism, teamwork, decision making, diagnostic reasoning, certain procedures and management, with some being difficult to define. Other constructs that are specific and are easier to define are suturing skills, intubation and needle insertion (37). In assessment, the construct is not physiological, but more of an educational objective (153) and needs to be clearly defined in order to interpret and provide conclusive evidence that the ongoing process is properly mapping the construct and that all irrelevant information is eliminated (37). Evidence should be gathered from many different sources when trying to evaluate

construct validity (153). However, it should be noted that constructs are usually not isolated measurable qualities and are not independent of content and context, therefore, assessment should be inferred from behaviour (24, 37). To ensure validity, the inferences must sample over a sum of knowledge areas (24).

It should be noted that assessment will rarely, if ever, be a perfect fit with the construct even if the construct is well-defined (37). In regards to higher-cognitive function, one can be able to say what it is, only when all of the laws involving it are known; and as this is not fully understood one cannot know precisely what it means (153). Rather, what can be used is what is known about it through cognitive taxonomies (e.g., Bloom) and factors affecting it (through cognitive theories). A more practical example is the construct of reaching a diagnosis of pericarditis (specific construct) where a given scenario involves multiple components that when put together captures the broader construct (higher cognitive skills). In Figure 4.1, an example of pericarditis is illustrated in an ECG that factors in all necessary information and skills required to answer the item. These include basic knowledge of anatomy, interpretation skills and decision-making skills, familiarity with ECG patterns of this diagnosis and being able to read an ECG report, knowledge of the natural history and pathophysiology of pericarditis, pharmacology knowledge of contraindicated medications, integrating symptoms with age, and knowledge of consultation and appropriate management.

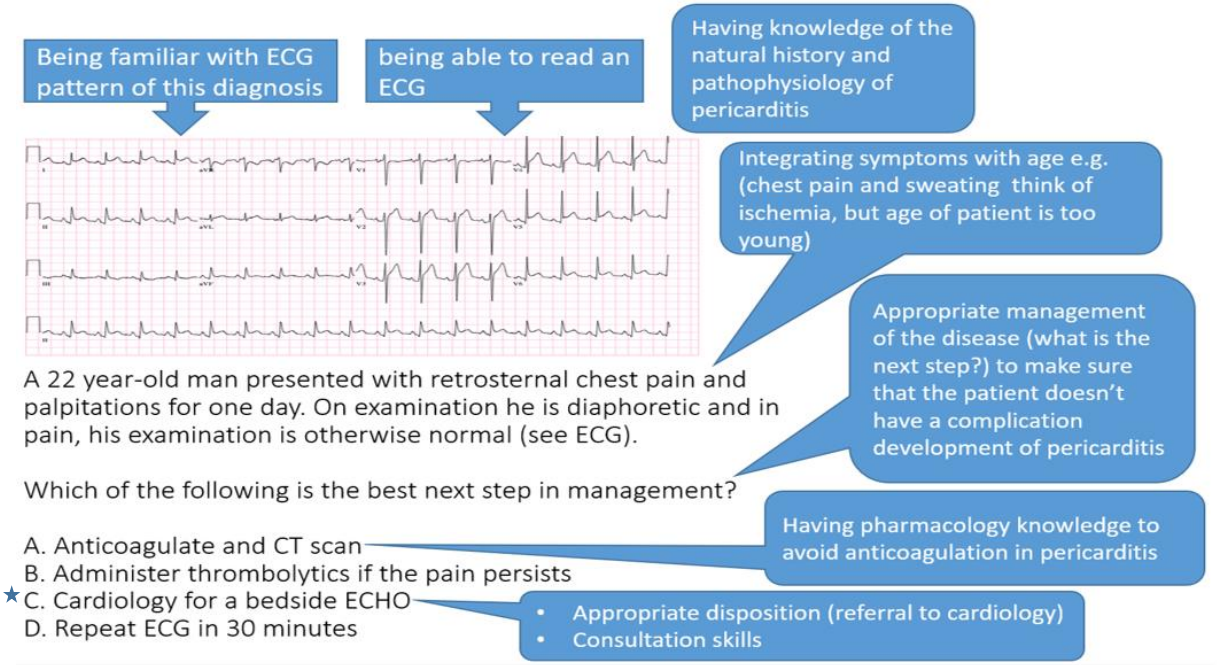


Figure 4.1: Practical example of a specific construct

There are two threats that can occur in any construct and that need to be kept in mind: assessment underrepresentation and assessment over-representation, and, of course, missing the construct completely (37, 154, 155). This is called construct-irrelevant test variance (CIV). CIV occurs when the test contains variances that are irrelevant due to something else other than the intended construct (154, 155). Assessment is considered limited when it is underrepresented. That is when it focuses, for example, on one aspect of the construct and does not reflect on the other aspect (or excludes it) which leads to the validity of decisions that are based on assessment results that are limited (37). Underrepresentation can occur if the exam content is too narrow and fails to include important aspects of the construct (154). For example, an EM exam that contains part of the curriculum content and is aimed at testing the construct of knowledge representation of the whole curriculum. On the other hand, assessment over-representation occurs when the assessment captures not only all the aspects of the construct under measurement, but also includes extraneous

aspects that are not part of the construct (156) “such as, in the case of EM knowledge assessment, including items related to research skills, language skills, speed etc.”. These additional aspects that have been included, even if important, are being tested by the learner and are not part of the construct that should be tested (155). This leads to a limitation of validity of judgments based on the produced results (37). A combination of both may also occur, where some aspects of the construct are missed (construct underrepresentation) and other irrelevant aspects are included (construct over-representation) (37). CIV can also result from test-wiseness due to item writing flaws that lead to scores being invalidly high (154).

4.3 Validity

Assessment is a central part of medical education for testing students’ capabilities; therefore, validity and validation are vital to its use. (157). Validity is an abstract contextual concept. It is essentially about making decisions from the interpretations and uses of test scores that are derived from assessment methods. Validity can be viewed as a task of identifying what the assessor is trying to measure (the construct), how they are attempting to measure it (the context and the tool), and how they use the interpreted data to make decisions and consequences (37). It is contextual in the sense that it is restricted to the context that it was applied to and does not fit all times, places, audiences, or purposes. It relies on the evidence gathered from examinees’ that have certain characteristics and applied in a certain context (37).

Messick described validity as ‘an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores’ (154 p.1). *The Standards for Educational and Psychological Testing* defines test validity as ‘the degree to which evidence and theory support the interpretation of test scores for

proposed uses of tests' (155 p.11). It can be concluded from the definition that: validity is not a property of a test, it is a property of the inferences made from the test scores; that scores from the same test may be valid for some purposes but not be valid for others; that scores are contextual because they are tied to a particular test; and that the construct and properties may not be generalizable to other situations (155).

Validity comprises of types and has a long history to it (158). Its meaning varies among different advocates for it and changes over time. It is the most sanctified lexicon in the educational and psychological testing community yet across its time there has been no common professional consensus over its meaning.

Validation, on the other hand, is the whole inquiry process that is undertaken to gather validity evidence to support the test defensibility of score interpretations (inferences). It is a cycle that one goes through over and over again and, therefore, it can never be fully attained (157, 159 p.171). Validation is a continuing process because validity is an evolving property (154, 157). Validation, according to *the Standards,* 'can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use' (155 p.11). The process of validating an inference can be simplified into two aspects" 1) developing a coherent argument for the proposed interpretation of assessment data (by gathering multiple evidence that is in agreement with the inference); and 2) Identifying arguments against probable alternative explanations (i.e., alternative inferences that are less well supported) (67, 154). These two steps have also been presented by different authors as 'a validity framework', which is later discussed. It should be noted that with all these definitions and frameworks regarding validity, what defines it is not the test scores themselves, rather it is the

appropriateness of the interpretations of the test scores and the inferences and actions that are made based on the test scores (AERA, APA, & NCME, 1999; Messick, 1989b) (159 p.170); and that validating an inference means that it is ascertained that the multiple evidenced gathered is in agreement with the inference and that the alternative inferences are less well supported (154).

4.3.1 What validity is not

Validity is not a property of the test; one cannot say that tests have reliabilities and validities. That means one cannot say that a test is valid or invalid. The properties are for the test responses and scores that are a function of the items, stimulus, tasks, examinee's response, and context (i.e., background, environment, setting) (37, 154, 159). Validity is not a stable characteristic of any given assessment, rather, it depends on the mode of collection, as well as the use of assessment scores and examinee characteristics. Therefore, validity judgments are limited to each application and context (37). That means the closer an assessment is to its original implementation (application and context), the more it can be argued that the validity evidence and results from the first context is applicable to the second, and vice versa (37). From this view, it can be concluded that when an assessment is described as being validated it does not mean that this assessment is appropriate for all purposes, times, audiences and all places (37). Validity does not have an all or none rule; therefore, it is important to understand that validity is a matter of degree (154) and that since the evidence collected is not always complete, validation is a matter of making the best practical case regarding the use of the test and what the test scores mean (154).

4.3.2 History

The field of validity assessment has been growing and evolving for more than 60 years (160). Over the past centuries, many individual scholars, thinkers, researchers, committees and professional organisations have tried to discuss, classify, think, re-think and find an end to the lack of consensus over validity. And while *the Standards* have made this point clear, yet no widespread definition is agreed upon by professionals (158). The topic of validity and its history is far greater than the scope of this research, however, the following points are a summary of highlights in validity throughout history that help shape what it has become today and how it has been developed into a framework.

During the 1920s, the word validity and its implications had officially found its way in the lexicon of educational and psychological testing; however, since then, it has been debated on what it means and what it should include and in recent years there seem to be no signs of this debate becoming any less. (158). In the 1950s, the qualities that needed to be specified before a test could be published and validated were not adequately conceptualized (153) and, therefore, the APA Committee on Psychological Tests differentiated four types of validity in order make clear recommendations. Thus, validity was broken into four types: content validity, predictive and concurrent criterion validity, and construct validity, which was later reduced in the 1960s to content, criterion-related and construct validities (154).

Each of these validities has a different emphasis on the criterion. For example, criterion-oriented validity procedures (predictive and concurrent) have a more central emphasis on the criterion behaviour rather than the type of behaviour occurring in the test, which is what content validity looks at. In construct validity, the trait or quality

underlying the test is what is important and not the scores and test behaviour (153). As validity only starts to get complicated, we can start by simplifying it into a table with categories as presented in Appendix 5. The 1966 edition of *The Standards* noted that these aspects of validity are not independent and that the study of a test would normally include all types of validity (APA, 1966 cited in (154)), with the 1974 edition including adverse impact, bias and other social consequences of test misuse to it (154). In the 1980s, a seminal paper by Samuel Messick had stirred up controversy regarding validity theory and what it should include. He stressed that validity could not be measured and needed to be inferred from evidence such as response processes (50). He also argued that validity should go beyond the traditional definition we know (i.e., achieving good measurement) and that it should also include consequences (i.e., consequential validity) (154, 158). However, Messick's work fused construct validity and consequential validity into one unified theory (161).

In addition, the 1985 *Standard* (APA, 1985 as cited in (154)) continued its view towards a unified view of validity with no longer referring to validity as types but as categories of validity evidence (e.g., content-related). In the newest edition of *The Standards* (APA, 2014), validity is not referred to as distinct types of validity; it is referred to as types of 'validity evidence' (155). They describe five sources of validity evidence, based on 1) test content; 2) response process; 3) internal structure; 4) relations to other variables; and 5) validity and consequence of testing (155). Construct validity was viewed as the main validity including all other types of validity. Yet, with this concept, the construct model seemed to make validation an endless process. In response, Kane (2013) although in agreement with Messick's position on construct validity being the central element and including consequential validity, went further to develop an argument-based approach to validation and identified what kinds of

evidence are required to gather in order to support test validation (161). Since then, different advocates have found ways to categorize validity into frameworks to make it more comprehensible and operational for educators and test users to use.

It is agreed in many articles that validity is the most important feature of a score-based inference, and perhaps the most important notion in measurement and test development (155, 158, 162). It is also agreed that validity can be viewed differently according to which perspective one looks through. In some instances, as for the classical definition, validity seems straightforward and logical with interpreted evidence either supporting or refuting the validity claim and measurement procedure. However, in other cases, validity is not straightforward or logical, with disagreements amongst scholars on what validity should include, such as reliability, consequence, and other evaluative concepts (158).

Newton and Shaw (2016) took on a unique view on validity and describe, in their paper titled ‘Disagreement over the best way to use the word ‘validity’ and options for reaching consensus’, five types of views to validity according to what it should include. They illustrate this transformation of validity meaning well through a series of definitions by various scholars and thinkers. As one reads through these definitions and explanations of validity, one can understand the different validity perspectives from a conservative, to a traditionalist to a liberal viewpoint (158). The five perspectives, as described by the authors, are labelled as traditionalist, liberal, conservative, ultraconservative and non-convergence and incompatibility perspectives (158), and to better understand these views Figure 4.2 was shaped.

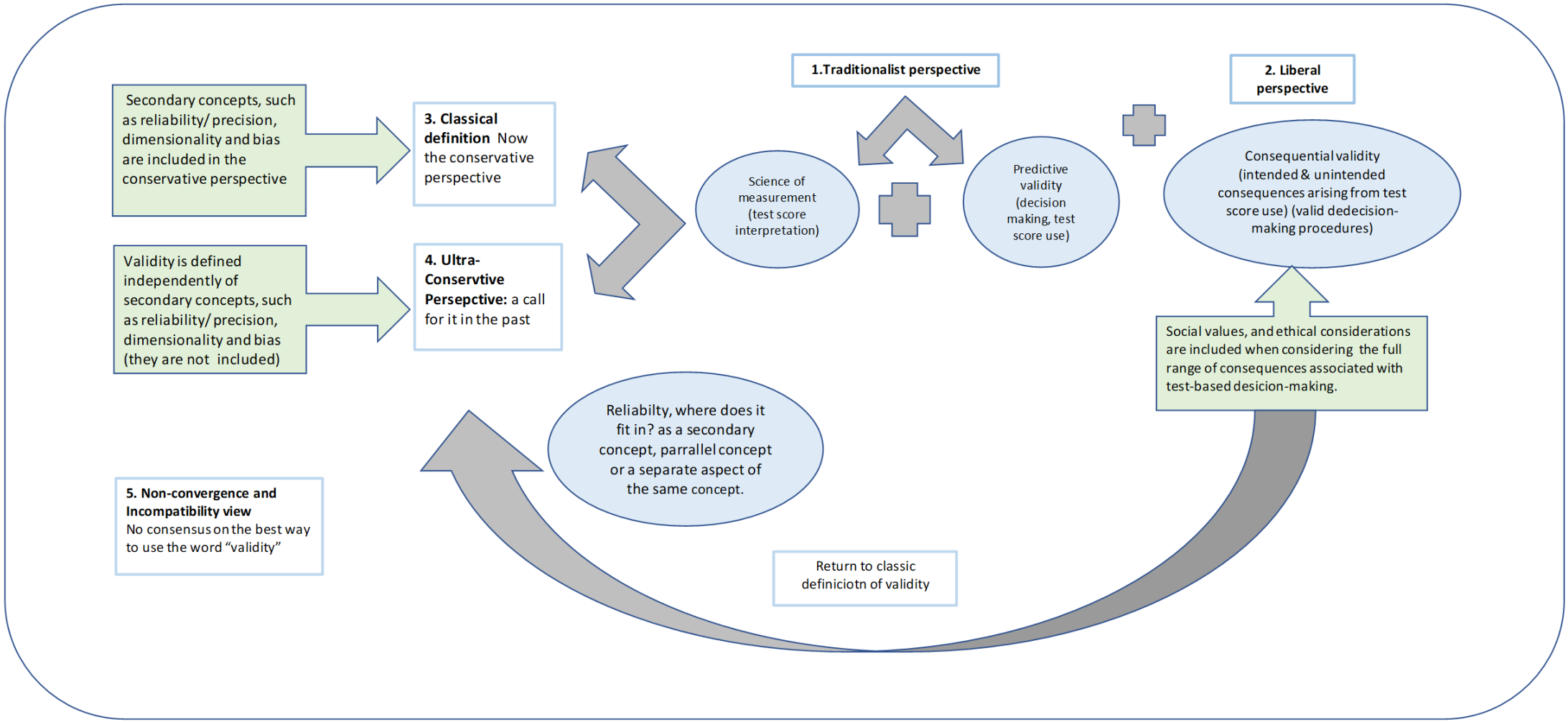


Figure 4.2: Validity perspectives as described by Newton and Shaw (2016)

From the illustration, the first view described was the traditionalist perspective of scholars. The traditionalist perspective of validity viewed validity in a 'technical or somewhat scientific' way as keeping up with the *Standards* and believed that validity should include both measurement and prediction, meaning that validity should focus on test score interpretation and test score use for decision-making, respectively (158). A traditionalist point of view to validity is inherently pragmatic. In other words, tests are created for a purpose, and the results are not generated to just simply be interpreted; they are generated to be used (158). However, even with this technical view, some tests were still recognised to be useful in certain conditions even if they had low predictive power. This outlines the idea of test usefulness. This usefulness considered aspects other than the technical view to embrace social, ethical, and economical concerns that are weighed against its costs and benefits (158). This led to the second view of scholars and thinkers that came to view in the last few decades as the liberal perspective of validity, and was coined as an extension of the traditionalist perspective as shown in Figure 4.2. The liberals' view argued the importance of consequences (intended and unintended) that arose from test score use to validity and validation and that it was irresponsible to evaluate a test only on the basis of test use and interpretation (i.e., the traditionalist perspective) (158). In recent years, a third new conservative perspective of validity has become prominent, returning to the classic definition of validity that it is "a degree to which a test measures what it is supposed to measure" (158). This view argued that validity should be viewed as a scientific concept and not a pragmatic one. In this perspective, valid prediction and uses of results or decision-making procedures are considered errors. This view is more narrowly focused on measurement quality such as reliability, precision, dimensionality and bias, which are considered as secondary concepts that are included within validity (158).

The fourth view, the ultra-conservatives, had a more narrow look on validity than the conservatives and in the past decade, they have argued that validity is what it's supposed to measure and that it was defined independently of the secondary concepts as reliability, which indicated that reliability was not included as part of validity (158). It should be noted that validity and reliability are the two most important psychometric properties in assessment procedures, and today, reliability is seen as one of the facets of validity (42). Reliability indicates the stability or consistency of test scores, which deals with the question 'will we get the same results consistently?' (1). Reliability implies how accurate test results are, by maximizing reproducibility and minimising measurement errors while validity is concerned with how accurate the test score interpretations are. It should also be noted that a test cannot be valid without reliability and that the presence of reliability is required for validity, however, it does not guarantee it (42, 159 p.163).This is true because every component of the testing process from item writing to scoring is a threat to reliability and, hence, validity.

The final view from Newton and Shaw's article was the non-convergence and incompatibility view, which claim that until now there is no consensus on what is the best way to use the word "validity" and that across the published resources what is apparent are divergent perspectives with advocates for the conservative perspective, as well as for the liberal perspective (158). The conservatives root for excluding consequences from validity because if included within the scope of validity, a negative test score use or bad decision-making might lead test users to annul or discard score interpretations. They also argue that the conception of validity is complicated enough and including ethical implication within validity will only make it unnecessarily more complicated and difficult for test users to understand and, hence, apply it (158).

On the other hand, the liberals root for including consequences because if excluded then a positive test score interpretation might lead test users to think that it is justifiable to use the test score. They also argue that the exclusion of ethical implication from validity might risk test evaluators of pardoning themselves from any responsibility for investigating adverse consequences.

The brief history of validity demonstrates how the thinking process and views regarding validity have been shaped and changed. Most discussions from the literature seem to centre on what validity should and should not include and has shifted from relying on content and criterion types of validity to content, criterion, and concurrent validity. Later, when all three forms have failed, construct validity would be called upon (154, 158, 159, 161). Ultimately, it seemed that all types of validity (content, criterion and construct), as well as reliability, were trying to measure the same thing “the target construct” (157), which led researchers and thinkers to favour a unified concept of validity, “construct validity”, as the whole of validity or “unitary validity” and discarding the other types (154, 159, 161).

4.3.3 Validity frameworks

The whole process of collecting and interpreting evidence to support decisions based on the intended interpretation of test scores is known as validation (157). A rigorous validation process involves detailing the claims and assumptions and reporting the interpretations and use arguments and when put together systematically are presented in a framework. Frameworks, as described by Pangaro and Ten Cate, ‘encompass a group of ideas or categories to reflect the educational goals against which a trainee’s level of competence or progress is gauged’ (163 p.1197). Different frameworks deliver different ways of looking at validity in examinations and have

different purposes (163). The purpose of any validity framework is to guide test developers in constructing their examination processes and gather evidence to support their claims regarding score interpretation. Frameworks demonstrate what is needed in a testing process (e.g., claims, evidence, methods), who is expected to use it (i.e., who is responsible in each step) and how to successfully apply it (i.e., clarity of classification and categories, ease of use, acceptability and value by users). In addition, how consistent, reliable and valid the framework can be applied (fairness) (163) and whether its applicability locally is as effective as its applicability internationally (generalizability). This will depend on to what extent it is understandable and what are the resources that are spent to carry out and train stakeholders for its use (163). Examples of frameworks that address claims and interpretive use arguments are the APA's framework in *The Standards* (155), Downing's framework (2003) that presented the combined conclusions of the American Educational Research Association (AERA), the APA and National Council on Measurement in Education (NCME) for validity evidences needed (164), the twelve steps for effective test development by Downing and Haladyna (2006) (165), Kane's framework (an argumentative-based approach to validation) (166), and the newly developed Cambridge framework (35).

4.3.3.1 The Standards

Validity in assessment has gone from a classical framework of validity of face, content, criterion and construct validity to a unified model where the evidence of validity is gathered systematically from five sources as mentioned earlier (155, 164). The evidence, as reported by the *Standards* (APA, 2014) is used to test a hypothesis against score interpretations on being valid for their intended use (160).

So, what does evidence of validity focus on? In all cases, validity evidence is construct driven (167), and focus on how well the assessment information and data explain, define, and outline the underlying construct in order to use the results to make effective and actual decisions about the construct (37).

The *Standards* (2014) developed by the AREA, APA and NCME provide a frame of references that one should address in the testing process (155). It comprises of overarching standards that guide users or developers with the primary focus of the standard and subsequent standards that are labelled to thematic clusters to further guide users. Appendix 6 outlines these overarching standards. The purpose of the *Standards* is to provide test takers, users, and publishers with criteria for the development and evaluation of tests, as well as to provide guidelines that help in the validation process (assess the validity of test score interpretation and intended uses) (155). The *Standards* also adopt five sources of validity evidence that focus on different aspects of the test development process (155). These sources of evidence, illuminate different aspects of validity and are used to evaluate the validity of the proposed interpretation of test scores for a given test (155) and are explained below:

4.3.3.1.1. Evidence based on test content

Analysing the relationship between test content and the construct provides validity evidence. Evidence of test content includes test questions, item format, wording, themes (domains and classifications), administration and scoring (37, 155). In addition, experts' opinions in the speciality play an important role, as they are knowledgeable about the target construct that is being measured and are able to produce relevant context and content for that speciality. Traditionally, this used to be known as 'face validity' in the classical validity framework (37). Evidence is collected

from test content to evaluate its appropriateness with the purpose of the test and the construct under study. However, if content validity is high, it still cannot measure a certain construct or skill alone. A careful review of the construct and test content help point to potential sources of construct underrepresentation or construct-irrelevance. (37, 155).

4.3.3.1.2 Evidence based on response processes

In addition to the content representing the underlying construct being measured, the cognitive and physical processes must also represent the construct (37). For example, the use of simulation for clinical training and examination is a more proper construct to represent the context of clinical practice than the use of MCQs (37). The theoretical and empirical analysis of test-takers' response processes can provide evidence regarding the fit between their performances and responses and the construct (here the cognitive process) being measured. Process validity evidence may come from test-takers' performances (e.g., item analysis, eye movement, and response time) of individual responses on the items, from different groups and from analysing relationships between tests and parts of the test to reveal differences in test score interpretations and construct meaning. In addition, evidence can be collected from empirical results that yield consistent results from tasks with similar processes, or contrasting results from tasks with different processes and experts' judgments in the field and their consistency in score interpretation and applying the appropriate criteria (37, 155).

4.3.3.1.3 Evidence based on internal structure

This relates to the relationships among test items and their interrelationships. It is, therefore, essential to study if the score accurately reflects the anticipated evaluation

of the construct, which would provide an accurate assessment of performance (37). Internal structure validity evidence is gathered through scoring criteria and algorithms, and the combination of data explained by experts in the field (37, 155). Evidence that a test implies a single dimension of behaviour (unidimensionality) is important and can be sought through item homogeneity. This can be drawn from item interrelationships (reliability), the number of items, and differences in responses to certain items among different groups with a similar overall ability (i.e., differential item functioning (DIF)). In the end, it is important that scores accurately reflect the construct under study (37, 155).

4.3.3.1.4 Evidence based on relations to other variables

Relationships between the assessment results and other variables are another source for validity evidence, by representing the relationships between the results and other variables with predicted associations to the construct. If predictions are not observed, then possible explanations should be sought (37). This evidence tries to explore the intended interpretation of test scores for a construct in relation to some other external variable. This can be demonstrated by consistencies between scores across groups, tasks, and settings (through using reliability statistics) or having assessment differentiating between examinees' performance based on experience (higher scores for experts and lower for novice learners). Evidence may also come from the studies of relationships and consistency (generalizability) to other forms of test that either measure the same construct (convergence) or a different one (discriminant) or to criteria that a test is expected to predict. An example may be that scores on MCQ tests might relate closely to short essay items but may not be closely related to OSCE as it measures more of the clinical skills of a construct compared to the cognitive knowledge skills of an MCQ. Another important issue is validity generalization as

described by the *Standards* “the degree to which validity evidence based on test-criterion relations can be generalised to a new situation without further study of validity in that new situation” (155 p.18).

4.3.3.1.5 Evidence for validity and consequences of testing

The final source for validity assessment is the consequences, which demonstrates if the decisions made based on the assessment result work. Evidence in this area can be gathered from monitoring the outcomes of decisions (success or error) based on the scores and evaluating the intended and unintended consequences of using and interpreting test scores (37, 155). An example of an unintended consequence of a test could be an increase in procedure errors that may have led to a negative outcome and that of an intended consequence of a test is to have an increase in learner motivation (37). Consequences of tests can immediately follow score interpretations by test developers by having, for example, a pass/fail decision for a test or even a job interview. The evidence here relies on the validation process, which means gathering evidence and evaluating the soundness and appropriateness of the proposed interpretations of test scores, as well as their intended use. This evidence can include assessment of fairness, bias issues, and consequences of actions taken after the decision has been made (for e.g., if a decision has been made based on an assessment that over- or underestimates the actual competence in the construct) (37, 155).

To summarize these sources, evidence gathered from the content and response processes provides information about the learner’s performance that is related to the underlying construct. Data from these sources will be interpreted as scores in order to facilitate decision making. Evidence from the internal structure makes a connection

between the score and structure of the construct (37) while evidence based on relations to other variables tries to make connections to other tests that either measure the same or different construct. Lastly, evidence for consequences of testing looks for the soundness of results and the consequences of using them (37, 155).

4.3.3.2 Downing framework and Downing and Haladyna's 12 steps for test development:

The organisations from Downing's summarized conclusions that were based on the AERA, APA, and NCME work that was just explained determined that evidence should be collected under the five headings in the *Standards*; content, response process, internal structure, relationship to other variables, and consequences (155, 164). In his article, Downing tries to explain construct validity in the context of medical education and presents sources of evidence in a different format by giving examples and outlining typical sources of validity evidence for performance and written examination (164). Appendix 7 outlines the sources of validity evidence as explained by Downing.

To simplify this abstract language of validity, validation and its gathered evidence, Figure 4.3 demonstrates the general idea of validity framework and the APA framework was used for illustration.

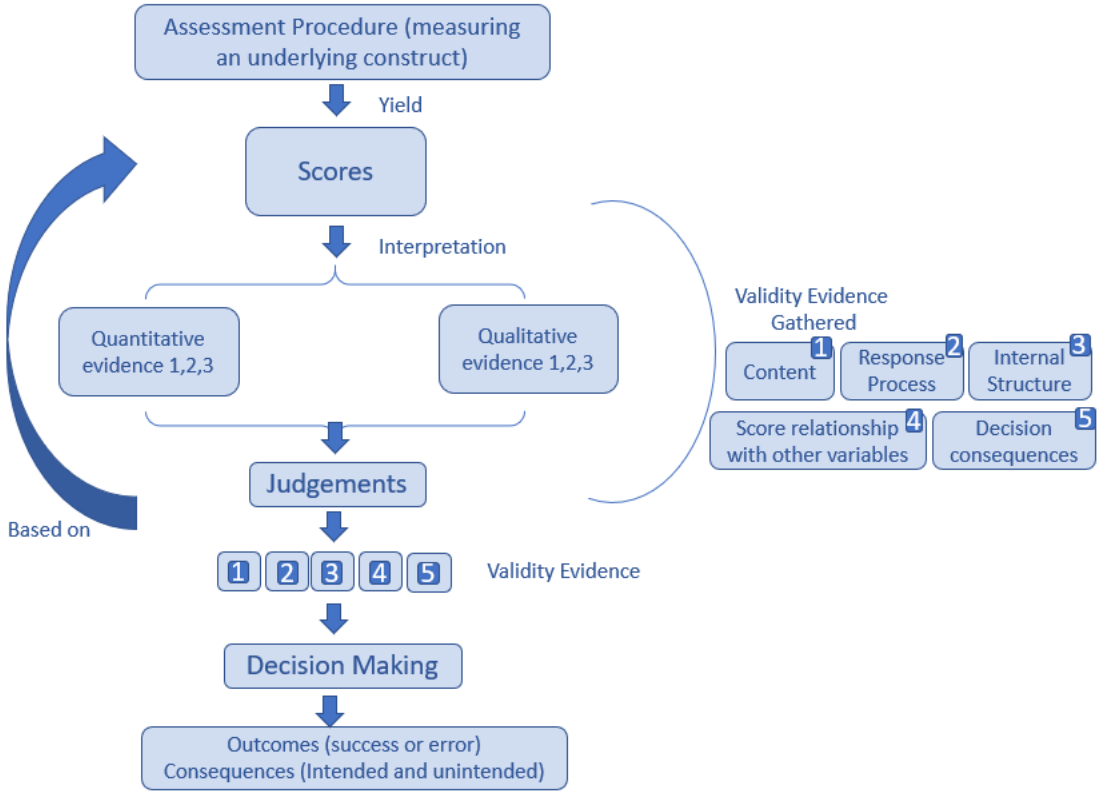


Figure 4.3: General concept of validity evidence and validation

Assessment procedures yield scores and judgments, and for us to say that these judgments are valid would require evidence. These “evidence” also known as “sources for validity evidence” are collected from methods and results of the assessment process and from theory to support and back up our actions and interpretations (inferences) that we make from test scores and are used for making decisions on the basis of assessment results (159 p.171). Based on the decision made from these results, an outcome of the test (either pass or fail) and consequences (either intended or unintended) arise. This figure simplifies the whole process of collecting evidence that is much lengthier and more complicated and can be better viewed in detail through the 12 components of test development (Appendix 8).

Downing with Haladyna recently presented this framework in a table, which makes this rigorous process more understandable and easier for the test user to apply (165, 168).

The 12 steps describe procedures or steps presented systematically as a framework that should be carried out in the development of most tests. These effective steps tend to maximize validity evidence for the intended interpretation of test scores and maximize the probability of measuring the construct under study effectively. In the first edition of *Handbook of Test Development* (2006) (165), these were labelled as steps, and a decade later in the second edition (2016), they were changed to components (168). Test developers find this template useful to go through these steps from overall planning and content definition to test production, administration, scoring and reporting. For each component (previously named steps), validity issues are highlighted and relevant *Standards* are referenced. The details of each procedure in each component depend on the purpose and type of the test. These are listed in a sequential timeline; however, they are interrelated and the orders can be modified with some being prerequisite to activities and others carried out simultaneously. Appendix 8 lists the twelve components for an effective test development process with their examples.

4.3.3.3 Kane's framework:

After gathering all the evidence, a validity argument needs to be structured by integrating various validity evidence into a coherent account demonstrating its support to the intended interpretation of test scores and its use (i.e., validation process). This is a never-ending iterative process, as there is always new evidence and information that can be gathered and used to support and understand tests and the inferences drawn from their scores.

Kane's approach to validity was an argument-based approach and his framework contained two kinds of arguments 1) the interpretation/use argument (IUA), that

specifies the claims to be evaluated (claims are the inferences in the proposed interpretation and uses of test scores); and 2) the validity argument that provided an evaluation for the proposed interpretation and uses of test scores.

The IUA relates to the intended interpretations of test scores and is used for the proposed context and population (166, 168, 169). It contains claims or, in other words, inferences under four sources of validity evidence (Scoring, Generalization, Extrapolation and Implication) (157) while the validity argument is the evaluation of the IUA through evidence that is collected throughout the test development process (166, 168, 169). Kane's approach is similar to that of the *Standards,* in that validity is presented as an argument-based approach but differs in its sources of evidence. Kane focuses on four key inferences, which are not embraced by the 2014 *Standards*. Instead, the five sources of evidence in the *Standards* are an emphasis on Messick's approach as explained by Cook, Brydges, Ginsburg and Hatala (2015) (157). Appendix 7 also contains the summary of Kane's framework and is aligned with Downing's proposal to get a better outlook on the differences between their outlines.

The final example of frameworks is the Cambridge framework, which is explained in the next chapter as one of the research methods that were outlined in Table 1.1 and was utilised in this research.

There are other theorists and researchers of validity with their own views of types on validity but what was presented is an overview of the most commonly cited literature.

4.4 Conclusion

An assessment task is not said to be ‘valid’ or invalid’, rather it varies by degrees and becomes either more or less convincing to the stakeholders (37). Validity frameworks address assessment concerns from test construction to final decision-making emphasizing key inferences in the assessment process. They are a structured approach that uses multiple methods and sources of gathering evidence drawn from inferences of different assessment outcomes, underpinned by the theoretical literature (35, 159 p.171). Evidence supporting these inferences are collected and evaluated to present a validity argument that ultimately facilitates the presentation of a defensible decision about examinees who were being assessed (157).Thus, test developers and users should understand that validity arguments for any assessment method concentrate its attention on how well it reflects on the constructs (37). They should also understand the strength and limitations of the assessment tool that is being used to make a decision in order to ensure that the judgements that are being made are sound (157). The following chapter presents some of this evidence to support the choice of quantitative and qualitative methods employed in this study to validate MCQ-MM testing of Emergency Medicine physicians in Saudi Arabia.

Chapter 5: Methods and Methodology

5.1 Introduction

This chapter covers the overall research approach, study type, paradigm and design, as well as the methods and the methodology (central approach) that were used to carry out the methods previously outlined in Table 1.1. This section also provides a justification for the mixed research methods that were used, the samples, recruitment process, how data collection and data analysis were undertaken throughout this research project. The methods section should always try to provide enough information for other researchers to be able to replicate the process through which one went through to collect their data (170). Therefore, the methods and processes drawn from both quantitative, as well as qualitative approaches were explained. Details of these processes were important to reflect and evaluate the validity of the multimedia items used, as well as the whole research process. They also represent sources of validity evidence for the Cambridge framework that was used in this research. The following section will first cover the research approach that deals with the research paradigm, sampling design, and phases of the research then goes on to cover the methods (quantitative and qualitative) that were used in this research, namely literature reviews (this was covered in chapters 2 and 3), questionnaire, pilot study, test, focus group, validity framework and legitimation.

5.2. Research approach

When conducting a research project, the researcher should consider which approach he/she would consider. There are two scientific methods that are used to carry out a research, a confirmatory approach and an exploratory approach (159 p.17). In the confirmatory approach (deductive method), the researcher starts from theory and states

a hypothesis that is based on an existing theory, collects the data to empirically test the hypothesis, and then either accepts or rejects the data. This approach uses the logic of justification. In the exploratory method (also known as the inductive method), the researcher starts by making an observation regarding the data, studies the observations for patterns and then makes a conclusion or generalization of the patterns to focus on theory discovery, generation, and construction. This method uses the logic of discovery (159 p.17,18). This research draws on both approaches and is more exploratory in nature as it tries to explore the use of multimedia MCQ examinations as a method of measuring higher constructs because little is known about its effect, as well as exploring the multimedia characteristics that one should take note of before being included in a high-stakes examination. The research also explores the use and applicability of the Cambridge Validity Framework as a means to evaluate the whole assessment process, to ensure that the validity of the test results and the inferences made from them are trustable. This research also has a confirmatory element to it as a hypothesis was stated (in Chapter 1) based on existing theory regarding multimedia and multimedia learning and after reviewing the literature.

After understanding the general approach underlying the research project, we can now try to select the appropriate study of this research and the type of research design that will help carry out this exploratory approach. According to Johnson and Christensen 2016, (159) research studies are placed on a continuum, with basic research at one end and applied research at the other end, and a research study can fall anywhere in between. This research lies in the middle of the continuum leaning more towards the applied research. Applied research focuses on topics that are the concerns of policymakers and

are driven by current problems in education (159 p1). This research aims to have a theoretical understanding of cognitive functioning level when solving multimedia multiple-choice items, thus reflecting a basic research element (159 p9) and aims at answering practical questions in assessment (Do MM-MCQs assess higher cognitive skills, and is the validity framework applicable for a high-stake examination in a new setting?). These questions aim to provide solutions that are valid and applicable, which reflects the applied element. The research study also aims to answer another important topic for policymakers and test developers, which is related to the concept of validity framework and how it is applied in the actual testing environment to ensure a valid interpretation of test results. Through applying and evaluating the validity framework, its worth, merits, and quality in the proposed examination, the process of application and evaluation helps to make a valued judgment regarding the use and appropriateness of the Cambridge framework and if its use should be continued or discontinued.

The research leans more towards the applied continuum because it is not applied in a strict laboratory condition as is in basic research studies. On the contrary, it is applied in a real-world setting (in this case the actual end-of-year EM examination, with actual application of the framework in the testing process). The aimed primary audiences for this research are other researchers, as well as exam policymakers, program directors and exam heads. All of these audiences continuously need to be updated through the literature on new policies and standards that can be applied in their assessment (159 p10). Another reason this research is more directed to the applied research spectrum is that this type of research often leads to an intervention or program development to improve societal conditions (in this case improving and changing examination strategies

to increase residents' competencies, in order to graduate safe and competent doctors for better patient care). Although it is out of the scope of this research to follow up on this chain of effect, the results of this research would help in taking the next type of research forward (i.e., implement these changes to improve licensing examination that screen overseas physicians) (159 p10).

5.2.1 Research paradigm

In research, there are three known research paradigms: quantitative, qualitative, and mixed-methods paradigms. The mixed-methods is an approach of using both qualitative and quantitative data in a single study (i.e., data collection, analysis and integration) that helps to inform results and provide a better understanding of what is being studied than if used alone (171, 172). The quantitative aspect of the data can present in different forms, such as tests (national, pre and post-tests), quantitized qualitative results, questionnaires, and multiple-choice assessment (171). while the qualitative aspect of the data can take the form of open-ended surveys, interviews, focus groups and video recordings (171).

The research paradigm taken in this study was the mixed-methods approach, which drew on both qualitative and quantitative paradigms. Therefore, it draws on both exploratory and confirmatory approaches as previously explained. It is based on the philosophy of pragmatism. Pragmatism is the philosophical position that what works in a particular situation, is what is important, justified or "valid" (159 p.32). The notion that researchers should take the either-or position (quantitative or qualitative) on research approach and not be able to mix both methods is known as the incompatibility thesis which started to be rejected in the 1990s when the pragmatic position was being favoured.

Pragmatists believe that both qualitative and quantitative types of research are very important and that the research questions should derive the methods. They advocate for integrating both qualitative and quantitative methods in a single study to utilize the strength of both methods when understanding and analysing the data (159, 173 p.32).

The pragmatic position also emphasizes that what is important is what justifies our problems, what works for us in a particular practice and a particular situation to be able to answer our own research question (159 p.32). The addition of qualitative study to quantitative study adds on a more holistic view of what is happening in the natural setting and gives a deeper dimension and layers to reality. There are many advantages and disadvantages of using a mixed-methods paradigm approach with the main points summarized in Appendix 9.

5.2.2 Research sampling design and phases

Because mixed-methods research uses both qualitative and quantitative sampling methods, therefore, the mixed sampling framework provided by Onwuegbuzie and Collins (174) was used for this research. According to their framework, the research design in mixed methods are classified against two criteria: time orientation of the components and relationship between qualitative and quantitative samples as explained in Table 5.1 (159, 174 p.275-6).

Table 5.1: Mixed-methods sampling framework

Time orientation criteria		
Concurrent	Data for the quantitative and qualitative phases are collected at the same time or roughly at the same time period and are then combined for interpretation at the interpretation stage	QUAN QUAL
Sequential	Data from the sample of one phase of the study (e.g., quantitative) are used to shape or structure the sample selection of the second phase (for e.g., the qualitative phase),	QUAN followed by QUAL
Sample orientation criteria (relation)		
Identical	Some individuals participate in both the quantitative and qualitative aspect of the study	
Parallel	Individuals in the quantitative and qualitative samples are different but are selected from the same population. A non-parallel sample is drawn from different populations.	
Nested	Individuals who were selected to be in one phase of the study represent a subset of those individuals who were selected for another part of the study.	
Multilevel	Samples of quantitative and qualitative individuals are taken from different levels of the population that is under study.	

Based on the above classifications, the sequential time orientation was used for time sampling, and the sampling relationship was identical for residents participating in both MCQs and focus group. Therefore, the design used in this research is a mixed-methods sequential design with an exploratory approach. Table 5.2 summarises the paradigm and approach taken in this research.

Table 5.2: Research approach, paradigm, epistemology and ontology

Research Approach scientific methods used to carry out a research	Exploratory approach
Paradigm Research model	Mixed-Methods (using qualitative and quantitative methods)
Philosophical Position The philosophy underpinning the approach	Pragmatism (what works for the research)
Sampling Framework Time orientation and relation	Sequential design (QUAN followed by QUAL)
Ontology What is reality, truth, and knowledge?	Reality is constantly debated, going back and forth listening to multiple forms of data and perspectives from residents and the literature
Epistemology Theory of knowledge (how reality is known)	Through finding out appropriate methods that solve problems, in this case, both qualitative and quantitative methods.

The research was conducted in three phases. The first and second phases of this research were concerned with quantitative data collection methods and results were derived from item analysis produced from the development of items for the pilot and main studies, as well as from the questionnaire. Information from these phases was further explored in the third phase of the research where a series of focus groups (qualitative data method) was conducted to probe for significant themes. The emerged themes were used to gain a better understating and explanation for the reasons behind the different statistical results between the MM-TXT matched questions. Hence, the design is viewed as a quantitative-qualitative sequential mixed-methods research design. Figure 5.1 Outlines the phases of the research taken and demonstrates the points of data collections. It also demonstrates where steps from the validity framework that were applied and its relation to the development of the examination process. The following section will elaborate more on the methods used in this research and the sample sizes for both qualitative and quantitative methods.

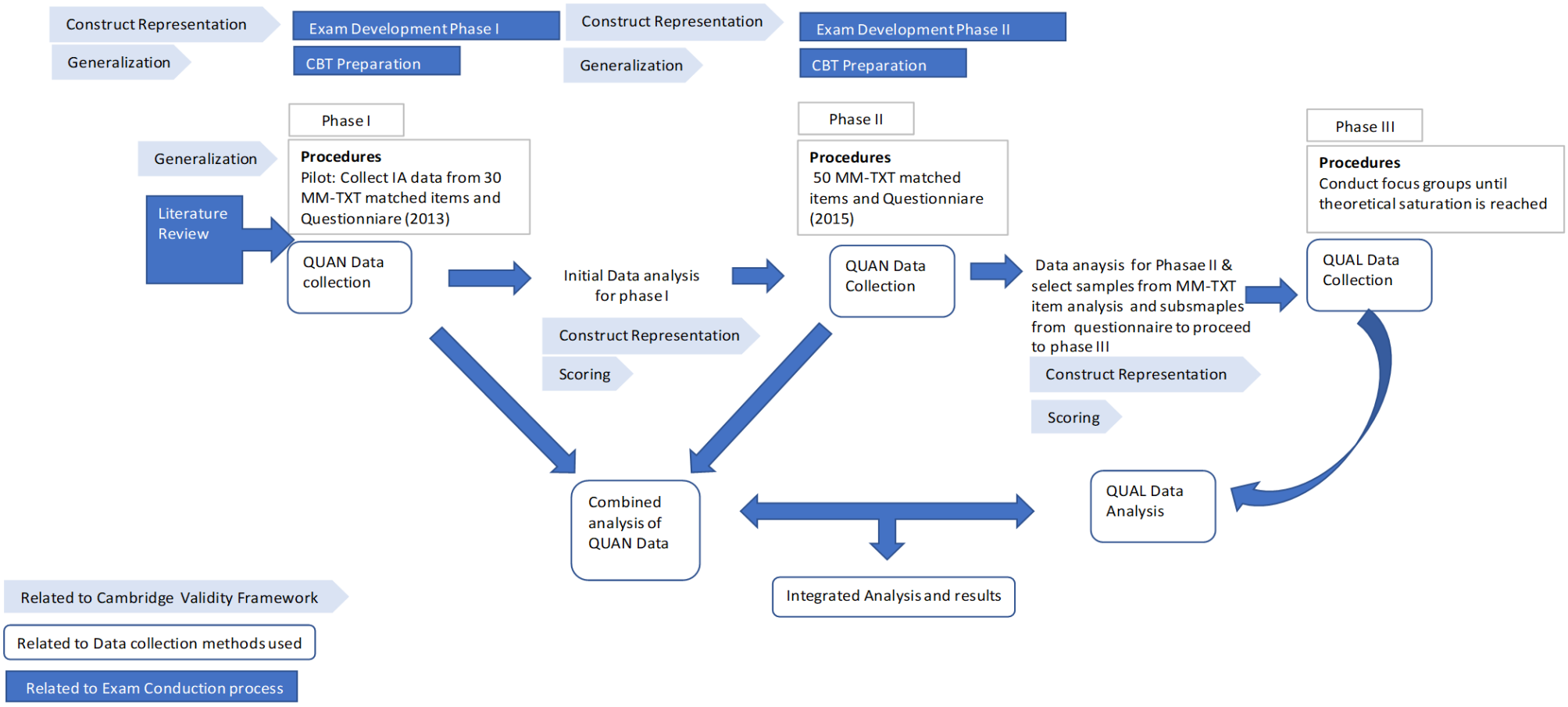


Figure 5.1: Phases of the mixed-methods research project

5.3. Research methods

The following methods that were listed in Table 1.1 were used in this mixed-methods research for data collection and are explained in more detail here (i.e., about the method, process, and analysis taken). Table 5.3 outlines the purpose of using each method in this research. As seen from the table below, data from the literature review, focus group and validity framework are qualitative in nature while data gathered from pilot and test item analysis and questionnaires are more quantitative in nature. The main purpose of using these methods together was to complement each other in order for one method to compensate for the weakness of the other, and to assist in drawing up a wider view of what the data and results represent.

Table: 5.3 Justification of the research methods used

Data collection method	Justification
Literature review (Qualitative)	To identify gaps in the literature, existing theories in multimedia that help in the conduction of the test. This method confirms findings and adds to the literature (175).
Pilot Test (Quantitative)	Informs decisions regarding the test and item appropriateness for test-takers. Provides information on the content of items Provides information on item statistics (159)
Test (MM-TXT) forms (Quantitative)	Describes residents' performance through statistical and item analysis and provides a reliable comparison between groups (171, 176) Increases the face validity of test questions (70). Provide the quantitative view of the picture
Questionnaire (Quantitative)	Provides overall view an quantitative information on how many participants hold a certain opinion (e.g., residents' experience towards CBT and MM type items) (177) Identifies sources of concerns to residents
Focus Group (Qualitative)	Describes residents' perspectives, experience, and thought process to fully understand the role of multimedia items in their examinations (171, 172). Identifies sources of confusion amongst residents Explores how opinions are constructed (177) Helps elaborate specific features from Quan results (171, 172) Supports Quan results (171). Complements findings from questionnaire and IA by probing for additional information and providing a qualitative view (171).
Validity framework (Qualitative)	Ensures quality for all aspects of the test process from start to end by following the framework in a systematic manner Supports both Quan and Qual methods by providing evidence of results' appropriateness (159).
Legitimation (Qualitative)	Ensures quality for all aspects of the research project from data collection to analysis. Supports Quan and Qual results through quality control and accuracy (159).

5.3.1 Literature review

One of the research methods that is used in all researches is the literature review and it is carried out for the purpose of identifying the researcher's topic if it has already been carried out. It helps in informing about the research questions, and understand what available methods and designs could be used to further help structure the study. Gaps in the literature can also be found where more areas are needed for research (159 p.86). Regarding the research topic, a comprehensive literature review was carried out and explained in Chapter 2 and discussed in Chapter 3 and the initial review demonstrated that the research topic still seemed to be at its infancy. The literature results were reviewed using the PICO strategy and PRISMA checklist as described in Chapter 2. Analysis of the literature and the theoretical background was carried out through a narrative review synthesising studies that were similar to the research and summarizing the results of the main findings (Appendix 2) (73). This helped shape the design of the multimedia items and understand what characteristics should be considered when designing them. As this helps to reduce the number of threats to validity and external factors that may affect test results and inferences made from them, hence affecting the validity of the study (155).

5.3.2 Questionnaires

Questionnaires are another common method for data collection and are considered to be a self-report data-collection instrument. It includes many questions and statements for participants to consider and respond to in order to allow the researcher to collect information about their behaviours, feelings, thoughts, perceptions, beliefs, and experiences (159 p.190). It is favoured amongst researchers because it is inexpensive, usually quick to complete by participants and a high response rate is not

uncommon to achieve. The problems with a questionnaire are that there is no probing, additional information and clarity cannot be achieved and it requires training in questionnaire development and delivery (159 p.228-9). One of the most important factors was to explore the successful implementation of multimedia in examination from the examinees' perspective and how they felt about it. Therefore, following the MCQ examination, students were given the choice of completing a questionnaire on his/ her overall experience with the format that they had received and to give their perspective on the overall experience with the examination's ease of use and relevance of content, to demonstrate the acceptability (face validity) of the examination

5.3.2.1 Questionnaire design

There are certain principles to follow when constructing a questionnaire that should be taken into consideration. The 15 principles listed by Johnson and Christensen were used as a guideline for developing the questionnaire (159 p.193). These can be found in Appendix 10 and were used as a guide for developing the research questionnaire. Since one of the objectives was to identify participants' perceptions, the questionnaire was designed to gain their insights and opinions regarding the use of multimedia items.

A self-completion questionnaire was developed drawing on points in the literature and using existing questionnaires available. Table 5.4 provides the list of terminologies used to search the literature. The questionnaire in this research was closed-ended and contained six broad themes, each containing a number of questions using a four-point Likert scale. The dimensions used in the categories were of 'Agreement' (1. Strongly Agree, 2. Agree, 3. Disagree, 4. Strongly Disagree). The categories in the questionnaire were selected to be a four-point rating scale that is commonly used by survey research experts, with no centre category to be included (i.e., neutral). Omitting

the middle category in rating scales does not affect the overall pattern of results and also provides less ambiguous data as it forces participants to lean one way or another on the categories (159 p.202).

Table 5.4: Terminologies used to search for questionnaire development

Search terminologies that were used in aiding the development of the questionnaires
Questionnaire/survey for evaluating exam items
Questionnaire/ survey for evaluating exam questions
Validating exam items
Validating exam questions
Validating new exam questions/items
Validating beta questions/items
Validating sample questions/items
Validating trial questions/items
Validating test items
Exam feedback survey
Pilot test
Post-exam survey
CBT
Beta test questionnaire
Survey for evaluating exam items

Clear instructions were used in the questionnaire and a lead-in statement for each section was used to orient participants for each new section (159 p.209). There was one contingency item (filter question) that directed the participant to a different follow-up question (159 p.209) for those who took the multimedia group (Section 2.C in the questionnaire), see Appendix 11. Each section of the questionnaire appeared separately on the screen in CBT allowing for more white space (i.e., screen is less crowded with more easily read questions per page). Readable font size, as well as different font styles (underline, bold and change of colour), was used to emphasize different sections and to aid the flow of questions. At the end of the questionnaire, a section was left for open comments. The questionnaire ended with a closing statement (159 p.210).

The questionnaire was developed in three stages. First, there was a pre-pilot stage where two experts in medical education and one in EM reviewed the questionnaire for its acceptability. Second, eight senior EM residents provided their input and comments. A minimum of 5-10 people is usually needed to pilot a questionnaire (159 p.211). This aided in checking for the readability of the questions, any additional refinements that were needed as well as identifying any points of confusion. Third, the questionnaire was adjusted and was piloted in 2013 (Phase I in Figure 5.1). It was distributed manually to 80 EM residents in three different regions in the country at the end of their promotion exam. Residents seemed to be able to answer most of the questions without much difficulty. Although participation was voluntary, almost all residents (n=78) completed the questionnaire with a few leaving out some questions. Almost all giving positive results and commenting on the good experience they had regarding the exam format being computer-based. The few comments that were suggested were taken into consideration for the following exam. A few minor errors in the questionnaires were noted and fixed for the 2015 questionnaire that was given electronically to the residents in their promotion exam (Phase II). The aim of the questionnaire was to assess the acceptability and perception on CBT, explore residents' views about the use of the new type of format (multimedia items), their relevance, clarity, quality in their examination, as well as how well the multimedia items resembled real-life cases.

5.3.2.2 Questionnaire analysis

The data from the questionnaires were analysed (Phase III) at the level of exploratory data analysis (descriptive statistics transformed into numbers and bar charts with prominent results further interpreted in the text) (178). The questionnaire was also

analysed to gain descriptive statistics for each question as a whole set in addition to having the results in groups as themes. Residents' comments were pooled together and were used to identify key issues and themes that were further used in the focus group discussions (179).

5.3.3 Pilot study and tests

This section explains the process of test development from beginning to end. It involves the two phases explained earlier: Phase I which is the pilot study conduction that involved developing 30 MM-TXT items and delivered to EM residents in their end-of-year examination through CBT, and Phase II which involved 50 MM-TXT and also delivered to EM residents through CBT. The process of both the pilot study and the main test were similar and explained here together covers two main issues in testing (as illustrated in Figure 5.1). The first is the process of conducting CBT and the second is the process of test development. All this was under the umbrella of the validity framework, which is also explained later in this chapter. A brief explanation on the concept of piloting, as well as testing as methods used in research will first be explained; second, the sampling size and methods used for selecting the residents, as well as the items for the study are described; third, seeking the ethical approval; fourth, the process of item writing in test development that was carried out in the pilot and testing phase, and finally, the analysis related to test development which is known as item analysis.

5.3.3.1 Pilot study

A pilot study is another method that is considered a crucial step in the research process, yet few seem to report the details of their experience, process, and outcomes with most only reporting the methods and tools that were used. Those that have been

reported failed to mention the details of what was learned and the changes that were undertaken. Reporting practical issues that one faces during the pilot set up would greatly assist others to avoid similar pitfalls and challenges. Because current research demands accountability, one must seek to use research results as best as possible. Some even take it further to argue that it is an ethical obligation to report these pilot phases as they contribute to the research experience and results (180).

Pilot studies satisfy a range of important purposes and can help in providing valuable insights for other researchers who wish to follow in the same steps and use similar methods or instruments. Even failed pilots provide insights on outcomes and processes to be avoided or that didn't work well (180). This is important because pilot studies can be costly, time-consuming, sometimes even frustrating and are faced with unexpected and surprising problems that one had not anticipated but had to deal with. It is, therefore, better to have to face these problems before investing a great deal of effort, time, and money in the full project. Although pilot studies are usually done using a small number of samples, they do yield areas for improvement and give indications on missed aspects that need to be considered for implementation in the actual project. Well designed and constructed pilot studies can inform about how to go about the best research process and outcomes (180). In this project, tests were developed for an EM end-of-year examination and were piloted to residents using CBT. This can be seen as Phase I in Figure 5.1 and the process will be explained shortly.

5.3.3.2 Tests

Tests are one of the commonly used methods for data collection to measure the performance of participants (159). Some tests may already be available and others need to be generated if they are trying to measure specific constructs or a certain

problem type. Tests may be designed to measure cognitive or memory processes or any other construct and, therefore, its content and context need to be tailored to the test's purpose (159 p.226). In this study, MM-TXT matched MCQ tests were used as a method to test higher cognitive skills. The mode of test administration was through computer-based testing in order to be able to deliver the exams to a large group of residents throughout the Kingdom. In addition, it's the only appropriate platform to deliver multimedia items. Advantages and challenges of CBT were covered in Chapter 3. Results from testing can yield quantitative or qualitative data or both depending on the type and purpose of the test. In this project, the results of the test are in the form of numerals presented as scores and item parameters which needed to demonstrate the differences and characteristics of the MM and TXT items.

5.3.3.3 Participants sampling

The data was collected from Saudi emergency medicine (EM) residents, which is considered a medium-sized speciality with around 80-100 residents. Residents were from first to third year participated. Fourth-year residents were initially included but were exempted a few months before the exam due to new SCFHS regulations.

The speciality was chosen for a) containing a wide range of cases representing different specialities in theory and practice (visually oriented); b) physician's regular exposure and use of multimedia in daily practice, and c) an existing good 'buy-in' for the proposed research within this speciality in Saudi Arabia, and the specialty content is likely to lend itself well to the use of multimedia. This method of sampling was purposive sampling (i.e., certain characteristics were required for the research and were located).

It is recommended when the population of a study is 100 or less to take the whole population, as the larger the sample size the smaller the sampling error is. A sample

size table for various populations of size 10 to 500 million based on a 95% confidence interval listed in Johnson and Christensen (2016) recommended that if the population size is 80 then the recommended sample is 66, if the population is 100 then the recommended sample is 80 (159 p.271). The more the group under study is of similar characteristics (i.e., homogeneous), the smaller the sample size can be because less noise is present (159). In addition, random assignment helps in strengthening the interpreted results as random assignment involves taking a specific group of people (usually occurs in convenience or purposive sampling) and assigning them randomly to a group that is being studied (159 p.269). In this case, the residents were randomly assigned to either the multimedia group or the text group. Before, assigning them, residents were divided according to levels and regions and then were randomly assigned using a random number generator. This is to ensure that residents were equally distributed among forms. Random assignment is used to produce groups that are similar in all possible factors and, hence, would be comparable at the start of the study. Therefore, any differences that occur between the groups was due to chance and any differences that occurred after the intervention was due to the independent variable (i.e., multimedia item) because it was the only factor that was different (159 p.269).

5.3.3.4 Item sampling

Regarding the items, the number of selected items to include in tests depends on the purpose of the examination and amount of time available to be tested on. In examinations, it is rarely practical or desirable to have an examination that is more than three hours (32 p.128). The number of items also depends on the domain that is needed to be covered on a test blueprint, the homogeneity of the test, and type of score interpretation that is required. The desired goal in an examination is to have

enough questions included so that most examinees have time to attempt all of it at their own pace (32 p.129). The number of items used in the pilot study was 30 and in the main study were 50. Items were combined (explained in results Chapter 6) to gain a bigger sample size of 80. Pre-test sample sizes for various test development applications reveals that in a single form of an exam developed from the best available items, using classical test theory (explained later in section 5.2.3.7.1) when one-third of the items are used, one would need $\geq$ 50 to 80 items, which was met in this study (181 p.495). A total of three hours of testing time per form was allowed. The number of items and time were based on previously unpublished work in the Middle East, for testing time on students whom English was not their native language. It was also based on the pilot study.

5.3.3.5 Seeking ethical approval

The first step taken to start the project was gaining the necessary ethical approval to conduct the pilot study, as well as the whole research project. This research project aimed to start with a pilot study that would be conducted to explore the use of multimedia in high-stakes examination in the speciality of emergency medicine. The setup for this study was arranged with the Emergency Medicine Scientific and Local Supervisory Committee of the Saudi Commission for Health Specialties (SCFHS). Meetings with the chairman, program directors, as well as residents were arranged to explain the steps taken to carry out the research project and to gain their initial approval in participating in the project. After completing the research proposal, ethical consideration was sought from the Saudi Jurisdiction through the SCFHS Council. As the project was being conducted at the SCFHS, additional ethical approval was sought from an external body IRB Committee. This was initiated to ensure that there was no bias towards the project, that the steps taken in the proposed project were appropriate,

and that no ethical violation was committed. The process was iterative and lengthy going through the proper channel of the hospital's IRP committee.

After gaining external approval permission, formal consent and agreement needed to be sought from the EM chairman and program directors for their residents to take part in the research and in answering the unmarked multimedia text-matched questions. Residents, however, would not know which questions would be the unmarked ones. It is common practice during examinations to include unmarked scores to obtain information about the items (pre-test beta items). This is practised in USMLE examination as well (20). The SCFHS rules and regulations have also included this. Unmarked questions have been used before in the Commissions' high-stake examination and were increasingly being implemented in different specialities. This process of adding these unmarked items was considered part of the normal assessment practice, as the introduction of unmarked questions were inserted into high-stakes examinations for piloting new questions or any new question format. Post-exam psychometric analysis of these items was reviewed to collect data on the performance and usefulness of those questions for future examinations.

The cooperation and collaboration of the EM Scientific Committee were vital not only in aiding the research process but also in finding new methods to improve the examination standards, which would help pave the way for other specialities and for the Saudi licensing examinations. The information sheet and consent form were created and distributed to the residents (see Appendix 12).

The following sections describe the item writing process to develop the pilot and test items from finding and training item writers to constructing the items according to appropriate item guidelines.

5.3.3.6 Item-writing process

The methods of testing and piloting should not be thought of as a single step. On the contrary, test development is more of a process involving multiple steps that require time and resources and need to be carried out properly in order to yield proper interpretations of test results. It should be noted that there are two iterative intertwined cycles here: 1) the item writing process and 2) the test development process, and both need to follow certain criteria and guidelines. The item writing process where the life cycle of an item development to test publication is illustrated in Figure 5.2, and to simplify, it generally involves four components, writing items and using multimedia, blueprint development, computer-based testing, and item analysis as outlined in Table 5.5.

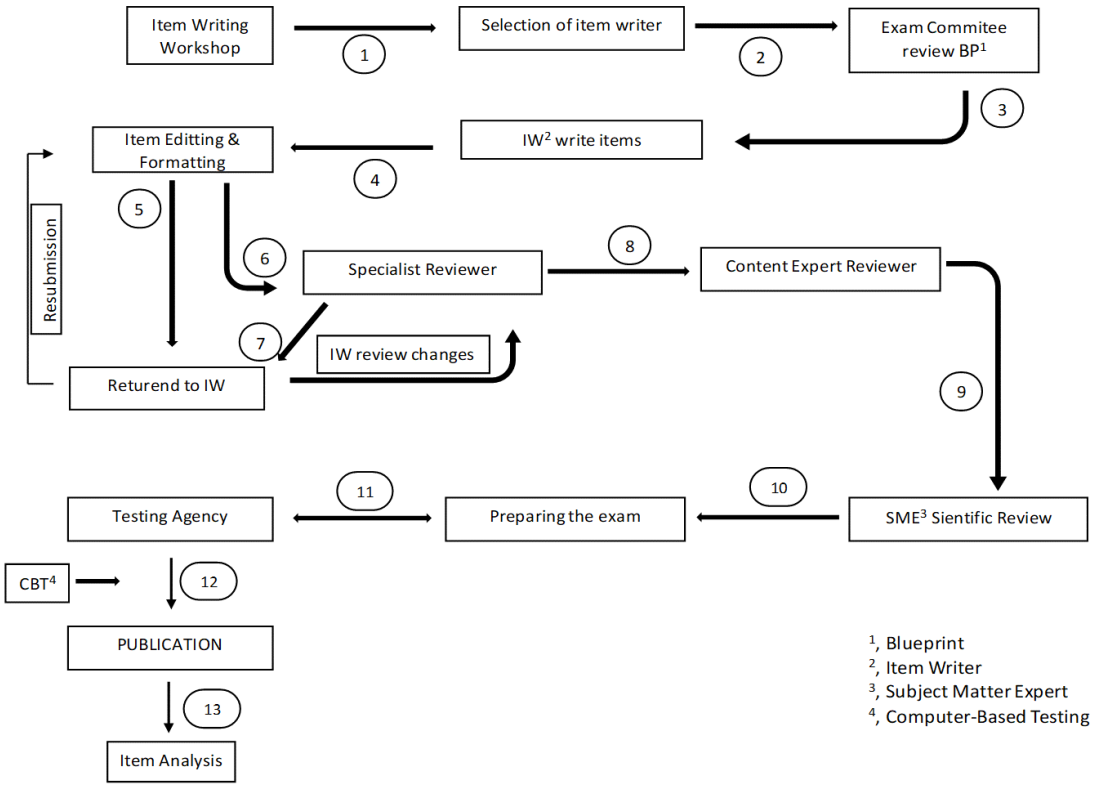


Figure 5.2 Item-writing process

While the test development process involves the item-writing process, as well as other steps as explained in Downing's and Haladyna's 12 steps for effective test development (168) (Appendix 8), it, therefore, is related to validity and the validity framework. In other words, the test development process involves the quality of the whole IW process and its components to gather evidence from item writing, blueprinting, CBT and item analysis to support or refute results' interpretations. To be clear to the reader, Downing's framework was used here to illustrate the difference between item and test development processes as it is outlined in a user-friendly way more than the other frameworks and would be easier for the assessment novice to comprehend. This is outlined in Table 5.5 by mapping the 12 steps of validity framework from Downing onto the item writing process components illustrated from Figure 5.2. The reporting and results of using the Cambridge framework as a method in this research is presented in Section 5.2.5.1 and its results in Chapter 6.

Table 5.5: Item-writing process and test development process

Item Writing Process Components	Guidelines	Test Development Process (Downing's Framework)
Item Writing and **Multimedia Use**	Item-writing guidelines Multimedia learning theory and cognitive load theory	**Item Development** **Test Design and Assembly** **Test Production** **Test Administration** **Test Security** **Test Documentation**
Blueprint Development	Guidelines for BP Cognitive taxonomy	**Domain Definition & Claims Statements** **Item Development** **Cut Scores** **Test Security** **Test Documentation**
Computers-Based Test	CBT guidelines and The *Standards*	**Content Specifications** **Test Design and Assembly** **Test Production** **Test Administration** **Test Score Reports** **Test Security** **Test Documentation**
Test Item Analysis	Psychometric analysis using CTT and generalizability theory to interpret the results.	**Content Specifications** **Scoring** **Cut Scores** **Test Score Reports** **Test Documentation**

The following sections explain the main issues related to the item-writing process (Figure 5.2) that were used for test development and that were applied to both the pilot study, as well as the main research study (Phases II and II in Figure 5.1).

5.3.3.6.1 Item-writing workshop and item writer selection

The first step in developing test items is to find and train the item writers who will develop them (step 1 and 2 in Figure 5.2). The literature is evident with the concept of faculty development. It has demonstrated that faculty development significantly improves the quality of test items and that input from peer reviews adds value to test development (182). In line with international guidelines, over the past two years, item writer workshops and extensive training process have been undertaken to create local item writer expertise in Saudi Arabia. As shown, Figure 5.2 demonstrates the item-writing process taken in SCFHS. Currently, around 26 emergency physicians have enrolled in SCFHS training workshops and assessment processes that were developed by the Medical Education Department, and 18 of them have been accepted as item writers and reviewers. Items arising from these workshops, and from newly validated writers have been reviewed and evaluated by the Commission's examination review committee. This final process leads to admission, rejection, or modification of items. Selected item writers would then be the primary authors of the new multimedia and paired written questions for the pilot and test study. The researcher with a team of medical educators was involved in training the item writers and reviewing their items.

In order to conduct the items necessary for this project, several steps had to be taken beforehand. The first thing required was to set up a revised and improved item writing workshop and recruiting item writers for the project. These workshops, which have been one of the newly implemented workshops, took participants through a series of

topics that included assessment, constructing high-quality items, item format theory, item writing flaws (IWF), as well as the use of multimedia. The second issue was recruiting the appropriate content experts to attend the workshop and write items. Therefore, e-mails were sent out to recommended emergency physicians who were active in the field of education and academia, as well as to a list of physicians from the SCFHS database.

Organisation of speakers, topics, and possible dates were arranged and were followed-up for appropriateness. Speakers selected for the workshop were medically qualified medical educationists (general surgeon, GP, and emergency medicine) with one being a senior in medical education and assessment for more than 20 years. The duration of the workshop was one full day. A presentation of SCFHS in-house style with an emphasis on writing higher-level questions was presented (step 1 in Figure 5.2). This was followed by several interactive sessions for review and critique. Topics that were chosen and updated with examples for the speciality of EM to be presented were as follows:

- Basics of assessment
- Blueprint and classification issues
- Recall vs. application (recall to reasoning exercise)
- Understanding performance data
- MCQ format, rules, images, multimedia and much more
- Balancing options, clarity and language
- Item writing flaws (IWF explained)
- SCFHS test development process

All topics were accompanied by examples, exercises, interactive sessions, and peer reviews. As suggested by the literature, all these are factors for improving the quality

of test item development through workshops (2, 182, 183). Presentations, exercise material, and references were printed out for participants to take notes and participate during the sessions. Item writers were also given an item writing manual, formatting general guidelines, a compressive quality checklist to be applied for each item, a normal value table based on SI units, and a structured template to insert their questions in. These were to ensure that item writers would produce questions that had the same layout and, therefore, ensure standardization of test format. After the workshop, an e-mail was sent with a zipped file containing the following materials:

1. Ten Item Submission (template)
2. SCHS Tables (vital and lab results)
3. SCFHS MCQ Manual
4. Item writing template instructions
5. Formatting General Instructions
6. Example MCQs

Participants were then required to submit 10 well-written items within two weeks according to the SCFHS in-house style, format, and guidelines that were presented in the workshop. Submitted items underwent a thorough review process with detailed written feedback for each question by SCFHS medical educators (including the researcher). The items were then sent back to the participants either for initial approval or requiring another submission of questions considering the given feedback. Once approved, a statement of work was issued for writing items according to the test blueprint specification they were assigned to (step 2). Follow-up of item writers and reminders were sent frequently in order to ensure proper review and quality of the items. Once the items were written, they then went through an iterative review process for formatting, content and language editing, clarity and relevancy check, item writing

flaw check, and much more before they were finally approved (steps 4-10). The review process was conducted by a group of medical experts in the field of EM and medical educationalists. Appendix 13 highlights what was checked by specialist reviewers and content expert reviewers.

A second workshop was conducted called the reviewer workshop approximately a few weeks later and focused only on the review process. Participants were only the accepted item writers who had attended the first workshop and gone through the training process. This workshop was also a full-day hands-on workshop; therefore, all participants were required to bring their own laptops. Pre-prepared materials were prepared on flash drives that consisted of a quick checklist of item writing flaws, examples of the item in their original format and after undergoing the review process, guidelines for reviewers, and illustrative examples to work on. After the completion of the workshop processes, the next step was selecting a few of the approved item writers to review the final test blueprint with the EM exam committee (step 3).

5.3.3.6.2 Reviewing the test blueprint

Blueprints, also known as test specification refers to how many items are used in an exam, its classification, weight, as well as which content topics and cognitive domains are included (184 p.186). Blueprinting is a necessary step for a valid and reliable test. The *Standards for Educational and Psychological Testing* makes many references to the test blueprint and the importance of content definition (155). Blueprints should precisely outline the percentages of questions allocated to the different domains and sections according to the assigned cognitive levels. (2) A clear relationship reflected through the blueprint should exist between educational objectives and the quality of test items. The use of test blueprints that are designed to include content domains, as

well as the levels of cognitive functioning, ensures that these cognitive levels, critical thinking and, by a means of extension, clinical competency, can be evaluated when developing multiple-choice items based on the blueprints (22).

The EM promotion exam is a paper and pencil end-of-year high-stake examination that permits residents to move from one level to the other. This exam is considered a high-stakes exam and must be passed by any doctor that is enrolled in the Saudi EM program and wishes to practice in the Kingdom of Saudi Arabia. It is not a certifying exam nor is it a specialist “board” examination. Those who do not pass must repeat the year and retake the exam the following year. Passing the promotion exam is a pre-requisite to entering the board examination. Rules, regulations, and exam specifications for SCFHS examination and the blueprint were set beforehand by the exam committees and local supervisory committees. However, with the newly implemented Medical Education Department, the newly developed guidelines, standards, and recommendations needed to be updated and implemented. The specialty of emergency medicine was one of the first specialities to undergo these changes as they had demonstrated their willingness to improve their training and assessment processes. And according to the new set guidelines by the SCFHS, each year, the blueprint was revised and could be adjusted according to major changes or advances in the program.

The EM blueprint (BP) was reviewed by emergency physicians that were selected by members of their scientific committee. The blueprint committee included a member from the scientific committee, EM item writers, EM exam committee members and a medical educator (the researcher) from SCFHS. Members had to fulfil certain criteria (e.g., involved in academia, teaching) with the most recent having gone through the SCFHS item writing workshop process. This selection aided in the process of writing

the questions as the members were familiar with the theoretical and scientific background of item writing. It also helped put things into perspective, minimizing resistance to change when laying down the blueprint and making the necessary adjustments.

5.3.3.6.2.1 Item classification, cognitive level and scoring

Cognition, as described by Haladyna, is " the act or process of knowing something" (33 p.20). A cognitive process is a non-observable event, and through evaluating examinees' responses one can infer that the examinee has a degree of knowledge and skills (33 p.21), but cannot reveal which type of cognitive process was used to answer a test item (33 p.27). In any examination, each MCQ has its own content and intended cognitive process attached to it that is intended to be measured (33 p.25); (27 p.13). However, no one can really know the exact cognitive process that is taking place when answering these items, and no statistical test can actually reflect this process. We can only approximate what the examinee is thinking when taking the items (33 p.25). This is not only as a result of this process being an invisible one but also because of how the memory works between an expert and a novice physician. When a complex item is written for a certain cognitive level, the expert working from memory and using a well-organised network of knowledge can simply respond to this complex item. Whereas, a novice physician will have to engage in a more complex strategy of problem-solving and a higher level of thought process to arrive at the same answer (33 p.26). This was also explained in Chapter 3 under cognitive schema (Section 3.3.7.3.4.8) and level of reader (Section 3.3.7.3.4.11).

To ensure content validity, item sets needed to be related to the learning objectives, as well as reflect teaching (23). The blueprint template was, therefore, revised by the

committee for content, sections required (e.g., cardiology, neurology, etc.) domains (diagnosis, management, pathophysiology and others), levels of learning from recall to problem solving (K1, K2-A, K2-B, K2-C) and classification (section, subsection or topic, section domain, system, item level, form and reference) (185). Table 5.6 explains the cognitive levels used in SCFHS and which were taken in this research. The cognitive taxonomy adopted (K1, K2) are based on Bloom's taxonomy that discriminates the different levels of cognition. K1 indicates comprehension and recall of memorized facts and is based on the lowest two levels of Bloom's taxonomy while K2 represents the higher levels of cognitive taxonomy (interpretation, analysis and evaluation). The simplicity of this taxonomy is due to the complexity of Bloom's taxonomy to be applied on items as experts tend to disagree on which level should be assigned to an item (186, 187). A classification system was added after each question, to ease for classifying the items into the bank**.** The classification system included: sections (EMS, toxicology, ethics, etc.)**,** subsections (chest, common paediatric problems), domains (diagnosis, management)**,** level (R1, R2, etc.), cognition (K1, K2), form (MM, TXT), reference and item ID.

Table 5.6: SCFHS item cognitive taxonomy

Item Level	Cognition Level
K2 (A)	Application, analysis and synthesis of clinical data with audio-visual
K2 (B)	Application, analysis and synthesis of clinical data with no audio-visual
K2 (C)	Recall and Understanding
K1 (D)	Isolated recall of fact

The SCFHS new examination rules aimed that at least 30% of the questions should be of recall type and 70% of the questions should be critical thinking questions. Although the literature doesn't specify the proportion of test items that should be written at the higher cognitive levels, it should reflect the practice, which requires a higher degree of cognition (2). Therefore, question difficulty was taken into

consideration when preparing the exam in order to make sure that questions addressed appropriate levels of residents (R1-R3) and that distribution of questions was appropriate for all (not too many senior or junior questions). The difficulty of an item to the examinees, in this case, will depend on the nature of the group taking the exam and whether they possess the required ability that is presented in the exam (32 p.131).

Although the EM program has objectives set for each course and year, the exam blueprint was based on the academic activity core content that is attended weekly by all residents. SCFHS promotion examination for most specialities is the same exam that is given to all residents with the only difference being in the cut scores assignment for each residency level (ranging from 55% for R_1 to 70% for R_4). Therefore, a balance was necessary when laying down the questions to ensure that the exam is neither too easy nor too hard or biased towards one residency level. Several meetings were held in order to revise the old EM blueprint and adjustments were made on some of the topics. Presented cases were those expected to occur in the healthcare settings within Saudi Arabia. Once the process of revision was completed, it was forwarded for revision and approval by the EM scientific committee. This was to ensure that the EM scientific committee was updated on new policies made in the examination through the medical education department and relay the changes to the training programs. Only the relevant blueprint sections were then sent out to selected item writers who had gone through the workshop process and were approved to write questions according to the domain and subject areas in the blueprint.

Scoring inferences is greatly dependent on the type of item used (e.g., wording of the item, task specification of procedures and training of standardised patients), as well as its response process (e.g., dichotomous responses, the weighting of response

options, global rating scales). The type of question and its response process also help shape the scoring rubric and item analysis (157). The promotion examination was marked only for correct answers as this is the current practice in SCFHS. Therefore, only dichotomous marking (0 for an incorrect answer, 1 for a correct answer) was assigned to each item for the computation of the psychometric properties. There was no negative marking for answering incorrectly. The items of the paired TXT-MM were calculated but were unmarked in the residents' total score and residents were informed of this during their orientation and a reminder e-mail was sent to them before the exam.

5.3.3.6.3 Developing the exam questions

This process applies to both the pilot and test study and consists of two parts: developing the unmarked multimedia text-matched questions for this research and developing and reviewing the marked promotion EM questions. In addition to the updated exam specification implemented by SCFHS on January 2013, additional exam specification needed to be applied for this EM speciality examination for two main reasons: 1) The exam would be transformed from paper and pencil-based to a computer-based examination, and would require further descriptions to ensure the quality and validity of the examination. 2) The inclusion of the unmarked beta multimedia text-matched questions into the examination. Therefore, all questions and detailed exam logistics needed to be delivered at least four months in advance of the scheduled exam date to a test administration industry that delivered all SCFHS computer-based examinations. An example of what a test specification might include is demonstrated in Appendix 14. It should be noted that not all exam specifications were identifiable at the beginning of the study; most had emerged during the pilot and development phase of the questions and preparing for the transition to a CBT format. However, it was vital that all specifications needed to be identified and stated clearly

at an early stage in order to be delivered and implemented by the test administration service.

5.3.3.6.3.1 Development and review process of promotion questions

Once the blueprint was allocated to each item writer to submit questions according to SCFHS item writing guidelines, an initial deadline was identified to be met for its completion. This allowed for sufficient time for feedback and review and to meet the CBT submission deadline. Items written were of the single best answer (SBA) MCQ format. SBA is one of three major MCQ formats that are increasingly being used. The other two being multiple true-false questions and extended matching questions (185). As described by Begum, 2012 'A Single best response or answer (SBA) - format consists of a list of possible answers, among which, only one is the "best" and the remaining are inferior but not incorrect.' SBA can be used to test the application of knowledge, problem-solving, and discrimination to a greater extent than the rest of the MCQ formats (185). The strength and weaknesses of MCQs were mentioned in the previous chapters and were collected from the literature and presented in Appendix 15.

A rigorous review process is essential for the development of a well-written question (2). Therefore, as soon as a set of exam questions were written, it went for initial formatting, reviewing editing by an assessment specialist and, finally, a review by a content expert specialist (steps 4-10 in Figure 5.2). The researcher was involved in the technical review and follow-up process of all the items. Once the review of the questions was completed, it was checked for any additional formatting. The questions were then sent back to the item writers with specific feedback and comments for

acceptance or rejection and review of the content. Questions were then gone for a final check by the EM program director and a member of the scientific committee.

In order to reduce the stress amongst residents that could be associated with the introduction of a new format of questions, as well as using a CBT format for the first time, a final review was conducted to label each question as either easy, medium, or difficult by EM consultants. This labelling assisted in arranging the sequence of questions in the exam and consideration was made to start the exam with easy questions and then gradually escalating to harder ones fluctuating in the middle (these were randomly arranged). It was also considered that the last few questions would be short and easy questions in case residents reached the end of the exam and needed to scan or review them quickly. Test item order can either be organised randomly, from easy to hard or from hard to easy. Studies show conflicting results regarding test item ordering and its effect on performance, with some studies showing that ordering items from easy to hard decreases test anxiety and increases student's performance and others showing that encountering easier items first makes the perception of difficult items even more challenging. On the other hand, some studies showed that tests performance can be compromised starting with hard items progressing to easier ones, which make students' feel that they do not possess the ability to perform well and others argue that students who started with harder items adapted, built confidence, and performed better (188).

In addition, the MM-TXT matched questions were electronically randomized within the promotion exam so that after every 3-4 promotion questions, a MM-TXT matched question in both forms would appear. This was done to ensure that the unmarked (beta) questions would be spread throughout the exam and not be focused in one section. Residents taking both forms would have the exact same sequence of

promotion and beta questions. However, during the exam, residents had the option of choosing from which question they would like to start from and could go back and forth in reviewing between questions as they wished.

5.3.3.6.3.2 Development and review process of paired (multimedia- text-matched) questions

In order to deliver multimedia questions, multimedia materials with certain criteria needed to be available to match the test blueprint and construct that was intended to be measured. First, a list of possible EM clinical problems with their diagnosis for questions that could be constructed in both the written description and multimedia format was generated (appendix 16). This list needed to a) be aligned with the test blueprint template; b) not duplicate the actual exam questions; c) not unbalance a certain section or domain of the blueprint; and d) be suitable for computer-based administration (45, 98). The goal in creating the list was to be able to a) identify a set of medical skills and processes that reflected higher order thinking and could be effectively measured through the use of MM enhanced MCQs; b) develop a prototype for MM items to assess the identified skills; and c) conduct a pilot using these items (98). The list was constructed and reviewed with two EM consultants and a surgical medical educator and was refined to include 80 possible topics. A brief description of what was expected of content in each topic was explained and prioritized. The list was later used as a guide to collect the appropriate MM materials from local and international physicians.

Second, a matched image or video that was copyrighted needed to be generated or found. The MM -TXT items are referred to as matched or paired items as they mirror each other in context and objective but differ in their presented format. One of the biggest faced challenges was finding copyrighted images and videos that would be

allowed for use. Communication with several sister-like organisations locally and internationally, websites, and physicians were carried out. Most of the process to acquire these images from other organisations was lengthy, timely and costly. A few websites replied with certain requirements that were not feasible. The majority of images and videos were collected from local and international credible EM physicians (124). Copyright permission to use the images for the purpose of this research was granted. Third, after attaining the copyright approval, the list of clinical problems that needed specific images was sent to these colleagues to see what was available in their image banks. Fourth, sent Images and videos needed to be reviewed for their quality, size, format of saving and appropriateness for the selected questions (e.g., age, gender, sensitivity, cultural issues, etc.). Adjustments were made to ensure proper image resolution, removal of any identifiable patient information if present and labelling images were completed. The clips had to be tailored to the type of information the residents should be tested on (122). Therefore, selected videos were cropped, adjusted, and edited to fit the exam format and were no longer than 10-20 seconds, as it is recommended that clips should be short (62, 122).

Finally, images and videos needed to be tested on SCFHS lab computers through a secure network on an agreed time and date by the testing agency. The secured network could only be accessed by machines pre-approved by the administrative agency as IP addresses were listed on their firewall to prevent unauthorized access. This was to review the quality and display of the multimedia questions on the screen and to ensure that both forms were present, correct, and that the function buttons were functional (i.e., enlarge, highlight, next, skip, etc.).

5.3.3.6.3.3 Item template development

After developing the items, their appearances and position on a CBT screen (item template) needed to be reviewed. Two versions of the test formats (pairs) were created in the domain of emergency medicine and each pair of items aimed at measuring identical content. Both versions were identical except for the medium (text or multimedia) used to present the clinical findings to the residents. In the multimedia versions, images and video were placed at the top of the question before the scenario. Instructions for the examinee were included to play and enlarge the multimedia.

Each question was linked to either a multimedia format (image or video) or text-based vignette presentation. Vignettes were written as a brief clinical introduction and would include patient's age, gender and chief complaint, related physical findings and if needed lab results and reports. This was followed by a closed-ended single best-answer question asking the examinees to identify the most appropriate management step or likely diagnosis. In total, 30 items were used in the pilot study and 50 in the main study.

Multimedia items were the first to be created and after being reviewed for accuracy and relevancy, parallel text-matched items were developed by describing the multimedia clips as text. Questions were reviewed by medically qualified medical educators (a surgeon, and three EM physicians) through separate arranged meetings to ensure that written and image questions were understood, similar and appropriate and was approved by the chairman of the exam committee. Each multimedia item needed to have been matched with the text item in their wordings, content coverage, clinical topic, as well as provide the same amount of clinical information (5). The exam committee also checked that the clarity, relevancy, and length of multimedia were

appropriate, as well as the difficulty level of the multimedia materials and questions (5). A sample of the multimedia questions was then given to a small group of recently graduated EM residents who represent the target examinee population for appropriateness. Modifications and refinements were made to be ready for the actual pilot testing. Figure 5.3 illustrates an example of a multimedia and text-match image used.

A 40 year-old lady injured her right hand with a kitchen knife. There is no active bleeding. On examination, there is a 1 cm laceration over the proximal phalanx of the right ring finger. Hand examination (see video).

Which tendon is most likely injured?

A. Flexor Indices
B. Flexor Carpi Ulnaris
C. Flexor Digitorum Profundus
D. Flexor Digitorum Superficialis

Key	C
Section	Trauma
Sub-section	Soft Tissue Injury
Section	Diagnosis
Item Level	K2-B
Difficulty	Medium
Form	B

A 40 year-old lady accidentally injured her right hand with a kitchen knife. There is no active bleeding. On Examination, there is a 1 CM laceration over the proximal phalanx of the right ring finger. On hand examination, the patient is unable to actively flex the distal phalanx when the middle phalanx is held in full extension by the examiner.

Which Tendon is most likely injured?

A. Flexor Indices
B. Flexor Carpi Ulnaris
C. Flexor Digitorum Profundus
D. Flexor Digitorum Superficialis

Key	C
Section	Trauma
Sub-section	Soft Tissue injury
Section	Diagnosis
Item Level	K2-C
Difficulty	Medium
Form	A

Figure 5.3: Example of a multimedia text-matched item

Item position was the same for both forms, and item position was arranged so that the paired items were interspersed throughout the test (every 3-4 exam item) and not clustered in one area. Students were randomly assigned to test forms using a

computer-generated system. Each student had the original 100 promotion exam questions in their exam in addition to 30 items (either multimedia or text) in 2013 and 50 items (multimedia or text) in 2015, depending on the format they were assigned to. The multimedia questions contained videos and images of ECGs, X-rays, CTs, ultrasounds, echocardiography and clinical pictures that would correspond to actual imaging methods used by emergency physicians in their clinical practice. No audio was included in this examination for two reasons: 1) audio testing was not part of the EM written examination objectives, as this is covered as part of their OSCE and clinical examinations and 2) the computer-lab was not equipped with headphones. All questions were written in the format of a single best answer, containing a single question and four response options (60). Each item was written to assess factual knowledge, understanding, interpretation or analysis so that the overall examination reflected the curriculum and level of residents, as well as being aligned with their blueprint (40). The cognitive load theory, as previously explained in Chapter 3, was used as the conceptual framework, using the principles of multimedia learning in designing the multimedia items (38).

5.3.3.6.4 Logistics for setting up a computer based examination

One of the most important aspects in setting up a CBT is proper communication and going through the appropriate channels with the various stakeholders involved who have different tasks, responsibilities, positions, backgrounds, languages and cultures. Several factors are involved and are intertwined and are dependent on the others for proper functioning. Involved stakeholders include item writers, item reviewers, formatters, test developers, secretaries, IT specialist, test centre administrator (TCA), exam coordinators, proctors, EM Committee members and EM program directors. Five important aspects of the exam needed to be completed before the date of the CBT: 1)

test environment and reservation of seats for candidates in the three testing regions; 2) preparing TCA's and exam materials in time for a review test; 3) completing all details of the test specification document (TSD); 4) preparing the residents; and 5) outlining test security issues and measures need for a fair examination delivery.

5.3.3.6.4.1 Test environment and seating arrangements

Seat reservation was essential as male and female examinees were tested in the same room but in separate stations and, therefore, the booking of enough seating for each region needed to be arranged and confirmed. Proper coordination was required at an early stage to secure the availability of the seats on the exact date of examination. This is because testing of other examinations was continuously running. Residents were assigned unique identification numbers (ID) that were used for specific seat assignment. The IDs were not known beyond the immediate research team. The central region (Riyadh) had the largest testing centre and, therefore, four stations were booked two for each gender. Because examinees might rotate in courses throughout their training, they needed to be notified beforehand on the available testing centres and which centre they would be assigned to. A total of eight stations were booked across the country.

The exam was administered in three main regions in Saudi Arabia, all in the computer-based testing centre in SCFHS-based headquarters. In 2013, 80 residents from R1 to R3 took the test while 84 residents took the exam in 2015. Once an examinee logged in, the software would display the random format that was pre-randomly assigned to them, either the multimedia or text form. Given that the examination was in SCFHS centres, examinees were provided with the same private work station, each equipped with a computer and a mouse. The work stations were

also equipped with head sound mufflers. Examinees were scheduled for a three-hour testing session with additional time allowed for viewing an orientation video and time to complete the questionnaire. On the day of the examination, the program director, as well as the head of the Examination Department and Medical Education Department were available to answer any questions or resolve any problems that may have arisen. Technical support was arranged by ensuring that a technical helpline was available and that IT specialists were on standby. Live communication with supervisors in the other regions was established through a WhatsApp group where any questions or changes were voiced and addressed.

5.3.3.6.4.2 Preparing TCAs and exam materials

Next, exam and TCA coordinators in all three regions needed to be familiarized with the exam process and characteristics. A face-to-face meeting was held with those in the central region and a teleconference session was held with those in the Western and Eastern province. A detailed explanation of the nature of the examination and what to expect was given. Materials were also sent to them with detailed instructions that needed to be followed in both languages: English and Arabic. Before the examination, a list of responsibilities for the day of the examination was also sent (e.g., proctoring, start and end time of examination, when to distribute questionnaires, what to do after the exam ends, etc.).

All exam materials and candidates' details needed to be arranged and delivered to the testing agency at an early stage in order to allow for proper implementation. These included:

- Residents' names, demographic information such as level and region.
- Residents' unique registration codes and form to be allocated to.

- Completed items (promotion and paired items) with their unique codes including the final images and multimedia files with their codes as well.
- Final template of both forms with the proper randomized sequence of questions
- An overall document that contained a list of items, specified form and which question contained multimedia
- Final blueprint for all questions (promotion and paired)
- Item bank count
- The final list of randomizations of forms to candidates, taking into consideration (level and region)

The registration process for candidates needed to be arranged. Residents were assigned a unique ID to this project only using the eligibility-based program to enable their tracking. This unique ID would link the candidate's name to the pre-assigned exam form. To protect the residents' privacy, no email, picture ID, or contact details were sent; however, on-site authentication was established.

5.3.3.6.4.3 Test specification document

A test specification document (TSD) or test specifications are specific test guidelines presented as a document that is outlined by the test developers to be given to test administers and publishers and involve all issues related to testing to deliver the test in a certain format (91, 155, 168). Preparing the test specification document was an extensive and iterative process that needed continuous review, sample viewing, and understanding the specific terminology and language involved. Before the TSD could be completed, a review of an existing exam was viewed and tested to be able to understand what the TSD was translating, and to see the usability of the exam. The review exam of a nursing speciality on a secure network was chosen for its similarity

in characteristics to the EM exam. The exam contained scenarios, long questions, calculators, videos, and images. The set up was initiated to see what was needed to be changed and implemented for the EM exam (e.g., layout and background of the exam, resolution and appearance of screen, feasibility of navigating the screen and using the buttons, presence of available functions (e.g., scrolling function), instruction and tutorial information, ending and beginning of exam, layout of review page, appearance and quality of images, videos and calculators, window size for viewing videos). After viewing the nursing review test, proper adjustments were recommended for the EM-TSD. Appendix 17 lays out more details and examples of what a TSD might include and were drawn from *The Standards,* the *Handbook of Test Development* and the ITC guidelines on computer-based and internet delivered testing (91, 155, 168).

After completing and signing the TSD, a review disc demonstrating the exam was delivered for any final adjustments or accidental mistakes. During the review exam, both forms were checked for correctness, the complete number of items per form, that all multimedia included within the correct question, all videos were clear and could be viewed and played, all images were clear and could be enlarged, and all calculators were inserted with the correct questions. After the first review disc, feedback and comments were sent to the testing agency to adjust some of the content, as well as image and video enlargement. The second review disc was later sent and approved. After finalizing the exam content, a date was set before the exam for a live demo to be viewed in all three testing centres. TCA coordinators in the Eastern and Western province were given a detailed checklist for the review process. All reserved computers in all testing centres were run to check for appropriate functioning. Also, the environment (e.g., seating, AC) was checked. Once the review test was approved, the test would be uploaded to be delivered on the scheduled exam date.

5.3.3.6.4.4 Preparing the residents

All EM residents participated in the multimedia text-matched questions as this was included with their promotion examinations as unmarked beta questions. This step (including beta questions) was methodologically important and essential for this project in order to be able to compare groups and draw conclusions from it. Therefore, supporting mechanisms were important and needed to be considered in order to minimise potential risks to the residents and to support the research itself (124). An orientation session was set up for residents on their academic half-day chaired by the EM program director and members of the scientific committee. A detailed presentation of the scope of the research project, as well as the Commission's implementation of a new innovative approach in their assessment process was explained. It was clarified that they were part of a research study programme, and were educated regarding the usage of multimedia questions in the CBT examination and that beta questions would be included with no marks assigned to it and that extra time to answer these items would be factored in to ensure that they had enough time to complete the examination.

They were also assured that their data and all their information would be anonymous and dealt with confidentially and that only relevant data for the research would be used. Afterwards, an explanatory and Q and A sessions were held by the program director and members of the scientific committee. Materials and examples of similar questions that would appear in the examination were demonstrated and were sent to all residents by the administrative assistants. Logistic materials containing (exam site and maps, exam contact information for venue, and general and CBT examination instructions in Arabic and English) were also sent. Questionnaire participation after the examination was explained to be voluntary

and would be at the end of the examination to minimise distraction and not to take up from their time. In addition, to the inclusion of an optional orientation video at the beginning of the examination—which was demonstrated as slides during the academic half-day—an additional instructional video was developed by the researcher and sent to all residents about a month before their examination. The video explained the process, format, layout and display of the examination. This helps reduce cognitive load as explained in Section 3.3.7.3.4.8.4 (Instructional Guidance).

5.3.3.6.4.5 Test security and measures

Test security is an important aspect of any test administration and assessment. It applies to both high and low-stake examinations, paper or computer-based. It is not an all or non-rule, it is a balance between the risk of a breach happening and the cost of preventing it (189). Test security, exams and other forms of assessment have increasingly grown in importance with the increasing role of testing around the world and with the aid of the rapid advancement in technology that assists in test administration, scoring and analysis (155). All stakeholders involved in the process of test development would agree that any form of cheating or test theft would diminish the value of that examination (189). Therefore, proctors, supervisors and test administers were all trained and certified by the commission and testing agency to be involved in the testing process. Policies and guidelines were put in place and stakeholders were reminded of them. Security measures related to the handling of the examination, as well as during proctoring of the examination was carried out. Items and multimedia were encrypted and sent separately to a secure server. Test-takers' pre-registration and onsite registration and authentication were implemented. Devices and mobile phones were collected prior to entering the examination. Residents' breaks were carefully managed and residents were escorted in and out by a proctor (TCA). A

single sheet of paper for note-taking was prepared beforehand on the station of the examinee and was collected afterwards from candidates before leaving the exam. Immediately after the exam, the data was stored in a remote server with strong encryption used during its transfer to SCFHS through a secure portal. All results, IA, reports, records, recordings and forms were stored in a safe place that only the researcher had access to. Electronic materials were encrypted, password-protected, and handled with care through a secure server and were stored password-protected on secured computers. All these measures, in addition to the security of item handling, contributed to delivering a fair high-stake examination that hopefully would not disadvantage candidates by allowing cheating or test fraud (189).

As a result of the high-stake of this examination, extra measures regarding security, logistical, and technical backup plans (e.g., backup paper print of the exam, IT standby, etc.) were considered in case of an exam crash. This was an important issue as the exam would be delivered only once and a retake was not possible. The system was set up so that in the case of an unfortunate circumstance where an exam crashed, it should be able to restart the exam from the point it crashed, saving all the data. It should also be able to build a result file for candidates who encountered technical issues. This served to be true after the completion of the pilot CBT examination, a failure of uploading of all files was because of a system crash. Although results were delayed for a few days, all data were retrieved and securely delivered. In addition, extra computers and seats were reserved in all regions in case of the occurrence of any technical issues (e.g., images not appearing, video not working, computer not functioning well). It was also arranged on the day of the examination that a help desk and helpline would be set up if any technical matter arose and that IT staff were on standby. A backup paper and pencil examination were also

available if needed. Any lost time in the examination not due to the candidate was factored in and given at the end of the examination.

5.3.3.7 Item analysis

Statistical properties of test items play an important role in identifying non-functioning items and are an important part of quality assurance as it aids in the improvement of test development (2). All items are routinely analysed after the examination for quality control and evaluation purposes (40, 190, 191), as illustrated in Figure 5.2 (step 13). This last section related to pilot and test study covers the analysis techniques used to analyse results from the data generated from the pilot and test study. An overview of classical test theory (CTT) and item parameters are explained, which were used as the main model for item analysis. In addition, reliability and G-theory were also used to compensate for any weakness from the CTT.

5.3.3.7.1 Classical test theory

Classical test theory was used in this study to calculate the test parameters using an in-house scoring item analysis software. In classical test theory (CTT), the observed test score is a function of the true score and random measurement error ($X = T + e$). In CTT, one can evaluate individual questions through item analysis and evaluate the overall test score. Typical statistic item parameters that are calculated in CTT are item difficulty, discrimination and distractor analysis (40, 192). Many previous studies used item analysis techniques that are derived from CTT because they are more understandable and are relatively easy to calculate. Other studies favour another technique called the item response theory (IRT) claiming that the collected information is superior to CTT. Statistical theories regarding which one to use for test scores (CTT or IRT) are still a matter of debate (33 p.203). The use of IRT in this study might have

resulted in different results if the sample size was larger. However, because IRT requires large sample sizes, it was not applicable for this data (49). While IRT may have better generalizable results, it does have its limitations, such as it needs a strong statistical background for analysis and it is difficult to calculate and needs to fit the model under study or results may be invalid. In addition, it has been shown that there is not much of a difference between CTT and IRT particularly for simple test items (49). CTT does have the limitation of quantifying error as being one whole error affecting a score where, in reality, it is multiple sources of errors. CTT deals with one error at a time (193). Therefore, in this study, this was compensated by calculating the G-Theory that allows for the calculation of multiple sources of error simultaneously (194).

Using the CTT, item statistics were generated for each item and for each item four indices were calculated. The first being item difficulty level (DIFF) also called P-value. Despite its name, it refers more to how easy an item is and is equal to the number of students who answered the item correctly (40, 192). SCFHS difficulty levels that are used are as follows: Easy is when >80% of the examinees get the item correct (DIFF =>0.8). Moderate difficulty is when around 45–70% of examinees get the answer correct (DIFF =0.45-0.79) and a difficult item is when < 30% of examinees answer the item correctly. A moderate difficult item is considered optimal, generally desirable, and yields the best scores for item discriminations. However, it can be varied according to test specification, format, purpose and level of the participants (40, 195). The second calculated indices was the index of item discrimination (DI) or item-total (biserial) correlations, which is calculated as the correlation between the item (wrong/right) and the total score. In other words, it discriminates between high and low-ability students and compares the response proportion of a correct item between the high and low-

ability students on the test as a whole (40, 192, 195). The third index was another measure of item discrimination and is the point biserial discrimination (rPB). It is the most direct method of correlation between item and test performance score (33, 40, 191) and is the most desirable one to be used (27 p.240). The DI and rPB are highly correlated and any one of them with a value below 0.20 are considered low (40, 191, 195). The fourth index was the examinee's response time to the items' "duration" and is calculated as the mean response time in seconds for each item. This psychometric analysis will also help in identifying item writing flaws and non-functioning items through determining item difficulty and discrimination indices (196).

Another parameter that is usually reported in testing is item bias. Item bias, as explained by JuHee, is "a measurement artefact at the level of an item". It may occur due to different reasons such as items not being included in a curriculum of a cultural group, the use of complex wording, cultural relevance, or inadequate translation. This type of bias is referred to as Differential Item Function (DIF) (197). DIF occurs when test takers have equal abilities on the content domain or construct but are from different groups (e.g., by gender, race or age) and they differ in their probability of answering an item correctly. The purpose of DIF is to identify construct-irrelevant variances (i.e., unexpected behaviour of items) in tests scores, format, content and scoring criteria and is commonly reported using IRT methods. It requires review and judgment and its occurrence doesn't always indicate a bias or unfairness in a test item. For example, if examinees from different levels have a different probability to give a certain response on an item because of their knowledge, it does not display DIF. Another example is if an item had different opinions amongst examinees and was felt to be culturally biased, it still can be retained in examinations if it is determined that the information it contains is important for safe and effective clinical practice, assuming

the information is available to all groups equally (198). However, it is not always feasible to measure bias for some subgroups as the number of members in the subgroup population or field test may limit the possibility of analysis. In these cases, focus groups or interviews may be conducted to search for evidence on the validity of interpretations that were made from test scores. Furthermore, sources of bias (CIV) and construct under representation should be prevented and looked for in a testing process (155).

In this research, IRT was not applicable because of the sample size restraint, as explained; therefore, DIF was not possible. Another reason DIF was not applicable is because residents were of the same culture and background (i.e., all Saudis). In addition, to ensure unbiased items in the examinations, experts were trained and were familiar with item construction, IWF, test content and target populations and reviewed the exam to ensure that items, format and stimuli did not contain information that was construct-irrelevant (such as language features that may be problematic, sensitive texts that may be sensitive to a particular group). Moreover, item writers and reviewers used simple vocabulary and language that were familiar to residents, at their level and comprehension and consistent with the purpose of the test without compromising the item's meaning, content and cognitive demand (22, 155, 184, 199-201 p.106). Reducing linguistic complexity (e.g., unnecessary, too complex, or peripheral wordings) does not affect the construct being measured; in fact, it reduces biases which, in turn, will reduce the threats to reliability and validity of test scores (200). This can be achieved through pre-test review and quality assessment (187, 202, 203). Finally, questions regarding item bias or difficulty were covered in the focus group discussions (155).

5.3.3.7.2 Reliability

Reliability, as previously defined, is concerned with how consistent the items are to each other and to the test. In this research, the reliability of the examination was determined by calculating two measures of correlation, Pearson correlation and Kuder-Richardson reliability, to measure the internal consistency of the items. Internal consistency is concerned with how steadily the items on a test measure a single construct, which indicates that the test is homogeneous or unidimensional (e.g., measuring knowledge) (159 p.164). A common rule of thumb is that coefficient alpha should be at a minimum, greater than or equal to 0.70 for research purposes and of greater value (> 90) for high-stake-stake examinations (159 p.164). Through maximizing reproducibility, reliability helps minimise measurement errors and is viewed as one of the facets of validity (42). Although all these statistical measures are important, the standard error of measurement (SEM) is considered one of the most important measures to report (204). The SEM signifies the differences between the observed and true scores that are associated with a particular test and represents the standard deviation of all errors of measurement in a test (155). The smaller the SEM, the less is the spread of the scores which indicates that the observed scores are more closely clustered around the true score (205). SEM is calculated by taking the square root of the error variance component (204)

5.3.3.7.3 Generalizability theory

Generalizability theory, also known as G-theory, is another measure of reliability and can differentiate between random and systematic error. Random error such as (guessing and noise) are beyond one's control and difficult to predict. However, systematic error constantly affects the examinees' scores because of test characteristics—for example, the difficulty level of an item due to the type of cognitive

level (194).

In any testing situation, there are threats (noise) to reliability that need to be considered and can contribute to the overall results individually or simultaneously (206). These threats are referred to as error variances (such as situations, items, observers and the interaction between them) and are based on the generalizability theory, which is an extension of the reliability in the CTT (193, 206). The components of these variances are analysed and are used to quantify the contribution of different sources of error and reflect the degree to which results measure the same construct. G-theory improves statistical power and can be used to calculate the reliability of combined samples and can estimate how many observations are needed (193, 206). It basically answers the central questions: to what extent can one extrapolate results from a test applied on a specific sample in a specific condition to a universe of conditions(193). G-theory has a range between 0 to 1 called the generalizability coefficient (G) and provides a measure of real differences that are detected between the examinees (206). An acceptable reliability threshold for high-stakes examinations would be G=0.8 (193, 206). Because this measure takes into account all the error sources at once it will have a lower value than the classical reliability coefficient (206).

G-theory is complex and most statisticians have limited experience with it (206). The G-coefficient has the ability in advance to specify the level of reliability that is necessary and calculate the number of items and situations needed by placing G in different hypothetical scenarios (206). This helps in judging the methodological quality of the assessment method (193). A D-study (dependability or modelling study) is the analysis that comes out from modelling G from pilot data and provides predictions for assessment situations in order to design an assessment with sufficient power (206, 207). It could be thought of as a statistical simulation. G-theory can also support

inferences of construct validity by testing the size of variance components (207). G-theory will allow the estimation of variances attributable to various factors that might influence variance in scores across items such as multimedia-or-written, as well as other factors such as candidate gender, region and level.

This concludes the long rigorous process of pilot and test methods used in this research that covered issues related to item writing, blueprint development, item development, and analysis and issues related to CBT, all of which are important to report as they represent sources of validity evidence. The following section is related to the fifth method outlined in Table 1.1 and is related to focus group conduction.

5.3.4 Focus groups

Focus groups (FG) are one of the well documented methodological tools for qualitative data collection (208, 209 p.16,32), are concerned with how participants think and make meaning from their experiences in the world (172), and are used in academic and medical education research (172, 208). They serve many purposes but are most commonly used for their in-depth exploration of a topic that we know little of, as they generate a large amount of qualitative data (208, 209 p.32). That means focus group discussions aim to try to understand an issue or situation and provide insights on how participants perceive issues or situations that they have experienced (208). A focus group is considered a type of interview where a discussion is moderated with a small group of individuals on their inputs and feelings about a certain topic (159 p 238). It is similar to an interview in that it allows the researcher to listen to participants talking, but differs in the sense that the talking is primarily done between participants rather than to the researcher, which is described as being more “naturalistic” (172, 177, 210). It is ‘focused’ in the sense that the participants are engaged and involved in a kind of

collective activity “i.e., the focus underpinning the discussion” such as debating and examining a particular set of questions (172, 211-213).

Focus groups are useful as complementary methods and are most commonly used in the exploratory phase of a research project (209 p.16). They are also used following quantitative phases of research to illuminate results through further explanations (209 p.45). Focus groups are widely used in combination with other methods of data collection (e.g., questionnaire) in the health sciences to enable researchers to enhance data finding through data triangulation (172, 177, 214) and are also used in mixed-methods research design as a means of triangulation, where different methods are applied in order to help in comparison and confirmation of results (172, 209 p.46).

Within this research, the focus groups were used to aid in the exploratory nature of the mixed-methods paradigm by providing insights on how participants perceived and thought of the different types of items. Therefore, focus group methods were triangulated with the questionnaire, and test results to add to the reliability of the study by revealing students’ perspectives from questionnaires with their perceptions from the group discussions and onto their performance on the test regarding question types (MM or TXT). This will add to the elaboration and clarification of both quantitative results of the item analysis and questionnaire (177, 212, 214).

The advantage of a focus group is in its inherent flexibility and their potential use in countless contexts (209 p.2); (159 p.239). Focus groups have the ability to facilitate comparison between groups that cannot be achieved by other methods (209 p.41). It is also used when one-to-one interviews are too difficult and are rendered time-consuming (209 p.42). One of its greatest capacity is to capture responses to events as they unfold (209 p.22). It is considered the method of choice when the purpose of

the research is to study group norms, processes, meanings, and decision-making processes, as it illuminates the inside ‘emic’ perspective of participants on how they made sense of the information provided and uncovered their inside misconceptions (209 p.33). Focus groups also provide a window to processes that may remain hidden or difficult to explore (209 p.26).

The organising and preparation of a focus group (methodology) require significant investment in time, resources and preparation from the researcher (215-217). From the literature, the methodology of conducting a focus group is explained as a guide for practical application on how one prepares and conducts a focus group (209, 215, 217, 218 p.2) and they usually fall under the headings of participant recruitment and selection, material preparation, session preparation, moderator role, group interactions and data analysis (172, 210, 215).

5.3.4.1 Participant recruitment and selection

Participants were EM residents recruited through the researcher’s personal networks, and through program directors. A list of residents was generated for those who took the multimedia items and those who took the text items and was given to program directors and chief residents to invite. All residents who took the multimedia (MM) or text (TXT) examination from the EM residents were categorized into MM or TXT group, this type of sampling, called purposive sampling, ensured that in each small group of participants, a homogeneous sample was selected for a more in-depth understanding about how the participants think. This type of sampling is commonly used in focus group recruitment (159, 172, 215 p.274-5), and is a type of non-random sampling. Non-random sampling includes convenience sampling (selecting participants who are easily accessible) (215) and purposive sampling (selecting participants who suit a

purpose in mind) (172). In this research, purposive sampling (deliberate selection of participants with a rich data source) was carried out. It is important that participants in the focus group share at least one important characteristic and are homogeneous in terms of background rather than attitudes (209 p.58); (159 p.238). The participants chosen were a group of Saudi emergency medicine residents who, within their groups, were acquainted with one another through their residency training program (shared a common background having the same profession) (219), although some of them did not work together regularly. Also, they all had undertaken the MM-TXT examination (shared a common focus) (219). So, they were considered to be fairly homogenous. Homogeneity of participants aids in providing a common ground for a cohesive group discussion, exchange of ideas, and allows for positive group dynamics (e.g., feeling safe when expressing concerns or conflicts) (172, 177, 215, 219). The disadvantage of homogeneity might be the lack of ideas and diversity within a group (172). However, in any given group, the participants are never entirely homogenous and would have a few differences between them to stimulate discussions (211, 213, 215). In this research, the residents' homogeneity was an advantage because friends and colleagues that shared daily lives, backgrounds, and experiences were able to have open discussions, encourage participation amongst each other, relate to each other's comments, and would often challenge each other on contradicting views on questions (172, 177, 211, 214, 215). In addition, the residents had differences between them, which was also important as they differed in their training level, gender, and type of questions received. These differences would reflect among them when they would disagree, misunderstand each other or ask each other questions to clarify why they think the way they do and explain their points of view (177, 211, 215).

5.3.4.2 Material preparation

It is important to prepare and make a list of all the logistics, equipment, and materials needed for the discussion. This includes providing a round table or oval seating arrangement, projector, a tape recorder, a microphone, notepad, as well as props such as flashcards or leaflets, spare audio-recorders and batteries (172, 176, 212, 215). In addition to these, the researcher prepared other required materials such as a discussion guideline, list of participants, consent forms, questionnaire results and a numbered sheet with of the questions that were to be discussed in the group for notes and commenting.

5.3.4.3 Session preparation

Session preparation involves decisions regarding how many focus groups need to be conducted, how long to run them, when and where to schedule them, what ethical considerations are needed and, finally, what needs to be known prior and during a discussion in order to run them.

5.3.4.3.1 Number and duration of sessions

There is no magic number for how many focus group sessions need to be conducted. Rather, it depends on how many comparisons the researcher wishes to conduct, the purpose of the research topic, and the type of data that needs to be gathered and analysed. In addition, the number of group sessions (sample size) depends on the complexity of the topic under study (172, 177, 212, 214) and the amount of new information that could be obtained from a new group (i.e., saturation) (172, 176, 215). In qualitative methods, a large enough sample selection is to achieve saturation. Saturation is reached when no new idea, information or relevant concept about the

research topic emerges from the different focus groups (172, 176, 214) (159 p.276). Nevertheless, holding two focus group sessions with similar characteristics makes claims about patterns that are found in the data firmer. This would suggest that the observed differences were not just a feature of one group (209 p.59). It is suggested that between four to six focus group sessions should be conducted to reach saturation and be able to generate adequate data and find patterns and themes (159, 172, 214, 215 p 238). Regarding the number of participants within each focus group session, it is said that four to ten participants are considered an ideal size (172, 177). However, a focus group can run perfectly well with 3-4 participants, and a group of eight participants could be quite challenging. The idea of moderating this number of participants is to be able to have group control, which allows for rich data to be generated without being overwhelmed when identifying voices and transcribing the data (172). The number of participants also depends on the size and layout of the room where the focus group would take place, as well as the complexity of the discussion desired (209 p.60); (159 p.238). In this research, five focus group sessions were conducted in three regions (Central, Western and Eastern) with a total of 33 residents that participated (a range of 4-9 participants in each group). Three sessions were conducted in the Central Region where the bulk of residents were, one in the Eastern Region, and the other in the Western Region.

The duration of a session usually lasts between one to three hours but can vary and is determined by the nature of the topic, as well as the number of participants in a group (159, 172, 177, 212, 215 p.239). This was apparent in this research between the first focus groups in comparison with the others as shown in Table 5.7. The conducted focus group sessions in this research lasted between 2-3 hours from start to end.

Table 5.7 Duration of focus group discussions

Focus Group	FG 1	FG 2	FG 3	FG 4	FG 5
Number of residents	4	5	7	8	9
Length of discussion (hours)	1.5	2.5	2.5	2.5	2

5.3.4.3.2 Setting of sessions

It is important to choose an appropriate time to schedule the sessions with residents and to take into consideration their other commitments (e.g., clinic, on calls) (212). The sessions were pre-scheduled with the program directors to be on a suitable date and time 'the residents' academic days' where the residents would be free of most commitments and would all be gathered together in the same place. It is also important to consider the site of the group discussion and provide an environment that is comfortable, has minimal distraction, and is most convenient for the participants (176, 177, 212, 215). Therefore, all meeting places were chosen to be in the residents' own environment, "the hospitals" where they worked in, and were arranged to be in a meeting/lecture rooms that had proper seating. Participants should be seated at a table and positioned in a way to have eye contact with the researcher and other members (172, 177, 212, 215, 219). Depending on the size of the room in each session, residents were seated either in a circle or U-shaped position to maximize face-to-face contact, and two recorders were placed in positions on the table to ensure that residents' spoken words were captured effectively from all seating areas (215).

5.3.4.3.3 Ethical consideration

As described in the literature, and covered in Section 5.2.3.5 (Seeking Ethical Approval), ethical consideration is required before the start of the study and participants should be contacted and provided with an information sheet and consent form about the study (172, 215). Prior to the discussion, the researcher should also

remind participants that the session would be audio-recorded and that note-taking would be employed (172, 212, 215) and he/she should stress the importance of confidentiality (209 p.67). Residents were clearly informed that there would be note-taking and recordings of the discussion and their agreement was obtained. The researcher also assured residents their confidentiality and anonymity regarding all collected information and results and provided them with an information sheet and informed consent to read and sign (Appendix 12).

5.3.4.3.4 Running the group discussion

Prior to the group session, the researcher prepared a guideline as a basis for discussion. This was to ensure that consistency was met across all focus groups and that the researcher was focused on the topic of study (172, 176, 209, 215 p.33), (159 p.239). The guideline consisted of the following: introduction (welcome, and explanation of the aim, format and sequence of the session), informal conversation, discussion questions starting with general ones (open-ended questions related to the questionnaire) and moving onto specific ones (MM-TXT matched questions selected based on their item analysis performance), and, finally, summary and ending of the session (172, 176, 212, 215). All sessions started with an introduction and explanation to the residents about the research, which was aimed at gaining an understanding of the participants' experience and thoughts into computer-based testing and multimedia items. Terms related to the focus groups were explained to participants (e.g., multimedia, text, item analysis, discrimination and difficulty index). This introduction is important to prepare participants for the discussion and ensure that they all have a clear understanding of what is coming in order to be able to contribute properly (176).

After the introduction and welcoming the residents, the participants were given around 15 minutes to interact informally and had the opportunity to address any concerns or specific issues they felt that they wanted to discuss. This ice-breaking session is vital as it enables participants and researchers to relax, get to be themselves, and set the atmosphere for the discussion (212, 215). After that, general open-ended questions were asked related to the questionnaire results, where residents were left to explain and answer according to their views. Materials such as pictures and posters can be used to augment questions as they stimulate discussion (215). This was true when 18 paired MM-TXT questions from the exams were displayed through PowerPoint slides to residents and the session became lively and active as they went through the questions. In fact, this is one of the advantages of focus groups—its ability to incorporate technology through the use of visual data. The visual data collection (i.e., images, photographs, cartoons, and videos) can be used to promote qualitative data that numbers alone can't communicate (159 p.245).

At the end of the session, key points were summarised and participants were debriefed allowing them time to raise any additional concerns that weren't covered in the discussion. Finally, they were thanked for their participation (209 p.95). In all five focus groups, the program director had popped in for a few minutes to encourage residents to express their views and ensured them that no harm or risk would be taken from their expressions. This was needed, as the researcher was from the SCFHS, the regulatory body for training and examining post-graduate specialities.

5.3.4.4 Moderator role

The moderator, sometimes called the facilitator of a discussion, takes a peripheral rather than a central role during the group discussion. This is because the most

important part of the group discussions is the interactions and dynamics of the participants. The moderator can be the researcher him/herself or an additional person to the researcher (172, 177, 208, 209, 212, 214 p.3). Therefore, the role of the researcher should aid in facilitating the discussion to encourage participants' lively and in-depth interaction by asking open-ended (prompt) questions, introductory comments 'openers', and follow-up/probe questions (172, 177, 208, 215) (159 p.232),(209 p.3, 83). Open-ended (prompt) questions facilitate discussions, provide clarification by asking participants to explain their comments, help synergise the contributions between participants to express their meanings, opinions, feelings and beliefs, as well as describe their individual experiences (172, 177, 208, 209, 215 p.83);(159 p.232). Openers (introductory comments) encourage discussions while probe questions help participants to relate to the initial question or to explore more (in-depth) on a topic (172, 215). Examples can be seen in Table 5.8 The researcher should spend most of the time probing participants' experiences asking them to compare and share their knowledge and discuss to what extent they agree or disagree with each other (172, 176, 177).

Table 5.8: Types of questions used in focus groups

Type of questions	**Examples**
Open-ended (prompt) questions	What do you mean? Anything else? Can you tell me what you are thinking? Why do you feel that way?
Probe (follow-up) questions	e.g., "Can you tell me more about that?" or "Can you give me an example?" Why did you say that the MM item was clearer? You disagree with her? So, what is your thought process in this question?
Introductory comments (openers)	"In the last group discussion, some participants felt that…"

The researcher as a moderator should be a good listener, probe for details, be able to pick up on different views amongst participants and explore them, encourage equal group participation, move the discussion forward when it is drifting and, ideally, should share some characteristics with the participants (e.g., age, gender, or language) (172, 209, 212, 215 p.4). The moderator may take a few notes but usually, the whole setting is captured through video or audio-recording so that the data can be analysed later (159, 172, 175, 215 p.239). Throughout this research, field notes were used, as well as a research logbook where ideas, discussions, challenges and concerns were recorded through-out this process (220 p.26). Jotted notes are not an uncommon use in focus group and interview discussions and include the condition of the session, place of the setting, who attended, and any interruptions that were faced. In addition, as an extra measure to the recorders that were being used, notes were taken about what was said, and if there was any particular expression that could be seen but not recorded (220 p.27).

5.3.4.5 Group interactions

The approach of focus groups has the potential for group interaction because it utilises the exchanges (dynamics) between participants as they talk and interact with each other (172, 177, 215, 221). Participants interact with each other as a group and also express individual opinions on parts of the discussions that they relate to. Focus group discussions offer the researcher an opportunity to observe group dynamics: agreements and disagreements, debates and challenges between participants, who dominates the discussion, who is shy or silent, and if someone changes their minds during the course of the discussion and shifts views (172, 177, 208, 214, 215). Depending on the topic, the discussion offers a chance to learn from the participants'

concerns, concepts, and language and offers a chance to see how participants engage in the process of sense-making. This includes what common assumptions are held, how their views are expressed (e.g., using different words for the same meaning) and how their thoughts are constructed and defended (210).

The researcher explored the differences in opinions between the residents and encouraged them to explain their views and theories as to why these differences existed. This process allowed for immediate clarification of what participants were saying and allowed them to reveal their underlying assumptions and tell their point of views in relation to other's perspectives (211, 212, 214, 215). Questions such as "What do you think? Do you agree with his/her point of view? What was your process of thinking when answering the question?" were asked by the researcher to encourage different members to speak.

During the sessions, the researcher also went around the group to give the opportunity for everyone to share their views. This is important to ensure that group members aren't being dominated by an individual and that not one person's view was being accepted as a group consensus (172). In the sessions, the researcher kept time, which is important to ensure that adequate time is left for the remainder of the questions as the discussion could get out of hand (176). Notes were taken by the researcher when residents had strong opinions on certain issues (e.g., time) and when they preferred a question to be presented in a certain way (multimedia or text). In addition, at the end of each discussion of paired items, the researcher would summarize the residents' points and thoughts on a prepared sheet to seek clarification and ensure what was said reflected their own views. This is important and useful to check out the groups' thoughts with the researcher's perceptions later on in the analysis phase (172)(209 p.112).

5.3.4.6 Analysing focus group data

There are various techniques that have been described in the literature for focus group data analysis, with no one method gaining worldwide approval (215). The analysis process of focus group data is the least developed and least agreed-upon part in the focus group process because it struggles with raw, transcribed information (212). However, the objective of the qualitative aspect of analysis is to understand the raw data and systematically organise and analyse it to determine the perspectives and assumptions from the participants' point of view. By reading and analysing the data, one gets a deeper understanding of what is being studied and is able to refine the interpretations throughout the whole process of analysis (217, 222).

The qualitative analysis phase cannot be considered a separate phase in itself in the research process, as it can start as soon as some data are available during the data collection phase and during the course of the fieldwork (217, 220, 222 p.27). Therefore, the first step taken in this process was writing up the field notes before any of it was forgotten, as this is considered one of the first steps in qualitative analysis (220 p.27). The strategy used for writing up field notes was inscription and transcription, meaning describing events and activities, as well as recording the participants' own dialogues and words (220 p.29). The analysis is a prolonged iterative process that involves movement between the whole and parts of the text and may take months to complete (170, 175, 212, 215, 217, 218, 222, 223). A comprehensive data analysis is commonly conducted after all focus group discussions have been completed (215).

Data analysis relies on both the researcher and the participants. The researcher is considered an instrument of the research process. Therefore, the researcher should have self-awareness of his/her thinking, personal perceptions, and biases when

analysing the data (215). The analysis also depends on the participants because the nature of the focus group data (discussions) depends on them. Consequently, the analysis is considered to be time and context-specific (208). The generated data (i.e., verbatim transcription) is at the level of both groups and individuals, which are usually difficult to disentangle from one another. It also contains different voices and speeches that would either be incomplete, not in agreement, or would be competing or interrupting each other (172, 177, 208, 215, 216). However, collectively, the emerged data comprises a number of expressions that captures the majority of the participants' perspectives and views that are supported by different proportions of the groups, in addition to individual opinions that are expressed (177, 208). The results of the data analysis would include a summary of the most important emerged themes with their noteworthy quotes to support it (176).

5.3.4.6.1 Understanding thematic analysis

There are diverse theoretical frameworks in qualitative research that help researchers decide which approach to select for data collection and analysis according to the purpose of their research and these include: building and generating plausible theories from the data (grounded theory); describing the significance of people's experiences and its meaning (phenomenology); understanding group culture with shared characteristics (ethnography); analysing linguistic expression (discourse analysis) and organising data systematically into a structured format (content and thematic analysis) (217, 218, 221). All these methods more or less overlap with thematic analysis because they all look for patterns and themes across the data set (217, 218). Grounded theory and phenomenology differ from thematic analysis in that they require to seek patterns in the data that are theoretically bounded (217). Thematic analysis is widely used and is often described as a flexible research tool that is used across other

approaches while, other times, it is considered as an approach on its own with only a couple of articles demonstrating guidelines for applying it methodologically (217, 218). Sometimes, thematic analysis is not explicitly described as a method of analysis on its own and Brain and Clarke, 2006 argued that a lot of the analysis is thematic but is either not identified or is claimed to be as something else (217).

In health sciences qualitative research, the literature demonstrates an overlap between content and thematic analysis with no clear agreement on how to apply it, and with content analysis often being the broad term used to describe analysis (170, 217). However, content analysis, although very similar to thematic analysis in summarising the descriptions of the data, usually incorporates a quantitative element to it (210, 217). This quantifying is usually to find significant meaning in the text but may risk removing meaning from the context and focus only on surface meaning (218). Thematic analysis, on the other hand, tends not to be quantified although, sometimes, some form of data transformation (transform qualitative data into a quantitative form) can be used if needed to describe or outline certain points (217).

In this research, thematic analysis was used as it provides a rich description of the entire data set and is appropriate for researchers with little or no experience with qualitative research. This is because it does not require detailed theoretical and technological knowledge of the other theoretical frameworks (217). Thematic analysis is a descriptive qualitative approach that researchers use to examine narrative materials, and break its text into small contents (units) to identify common or repeated patterns within the entire data set, analyse and interpret the patterns of meaning (themes), and then report it (217, 218, 224). A theme is sometimes referred as a category and, as described by Braun and Clarke (2006), “captures something important about the data in relation to the research question, and represents some

level of patterned response or meaning within the data set" (217). Polit and Hungler (1999) gave another description of a theme as 'a recurring regularity developed within categories or cutting across categories' (223). The researcher determines what constitutes a theme, as it is not necessarily dependant on the frequency or number of times it occurs in the text as it does in content analysis, but rather on whether it captures something important in relation to the research questions (217, 218).

It should be noted that there are two ways to analysing the data when using thematic analysis: an inductive way and a deductive way (172, 217, 218, 222). Inductive analysis (bottom-up way) is data-driven. This means that one searches for meanings and concepts when reading and coding from the data. The categories and themes are directly derived from the text and are, therefore, strongly linked to the data themselves and not trying to fit a pre-existing coding frame. It is usually used when little or no studies deal with the phenomenon under study (172, 217, 218, 222). While a deductive analysis (theoretical or top-down way) is analyst driven and is used when researchers want to apply a predetermined coding structure or set of themes (e.g., from the literature) to the transcript that is being analysed. Therefore, one looks for specific words and phrases in the data to support their research idea (172, 217, 218, 222). Its general aim is to test a previous theory in a new or different situation (218). In this research, coding was done through inductive reasoning with codes being defined from the transcript discussion and content of what residents have said.

5.3.4.6.2 Understanding codes

One of the most difficult parts of research analysis is its qualitative aspect (212, 222) of which coding is an important step taken. Coding is a dynamic intuitive process of inductive reasoning that involves reading the transcript, identifying passages that

represent a certain theme, looking for recurrent ideas, subdividing the textual data, organising it into meaningful groups to help make sense of it and then labelling it (170, 217, 220, 222 p.40). Initially, most identified codes overlap in meaning, which is further analysed to condense those that are related to each other into themes (170). A code is when one assigns a label to a unit of analysis (sentence or a collection of sentences with meaning) (176, 223). Codes help to think about the data in a different way and should be understood in relation to the context (223). Basit (2003) described codes as "tags or labels for allocating units of meaning to the descriptive or inferential information compiled during a study. Codes usually are attached to chunks of varying-sized words, phrases, sentences or whole paragraphs, connected or unconnected to a specific setting. They can take the form of a straightforward category label or a more complex one, for example, a metaphor (222)". The chunks of data are sometimes referred to as a unit of analysis. A unit of analysis (or meaning unit) is the basis for developing a coding system, which is then systematically applied across all transcripts (210, 223). As described by Braun and Clarke (2012), "codes are the smallest units of analysis that capture interesting features of the data (potentially) relevant to the research question. Codes are the building blocks for themes, (larger) patterns of meaning, underpinned by a central organizing concept- a shared core idea." (224). In this research 'sentences' were the unit of analysis and were a way to identify meaning to the segment of information in the verbatim text (175, 223). Each sentence could have one or more codes assigned to it from a list of codes that can be developed as the analysis proceeded. In addition, remarks and memos could be assigned to a coded sentence (175, 217). Table 5.9 outlines an example of how a unit of analysis (meaning unit) is transformed into code.

Table 5.9: Unit of analysis and coding

Examples of meaning units (sentences), edited meaning unit and codes	
Meaning Unit	R1: *I mean for me the* picture is not an integral part of the question, it's just something to add R-2: *This is correct and* I agree with you R1: *This is what we were thinking about that the* picture ok, is something to illustrate *but* I can answer the question from the stem alone without the picture without the audiovisual
Edited meaning unit	The picture was not integral to answer the question and was something to illustrate
Code	Acts as supplementary material (as one of the characteristics of multimedia)

When creating a code, we are making a decision about how the data needs to be organised in a way that is useful for the data analysis and that fits with the whole context. Therefore, creating codes is also a constructing process of developing conceptual schemes and trying to link between sets of concepts and ideas that are located in different areas in the data set (i.e., focus group discussions). This guides the researcher to compare data, ask questions, as well as create, change or drop codes to make themes and the appropriate data hierarchy (222). The most general broad theme in a hierarchy is referred to as the parent node with the child or children nodes stemming from them as branches. Codes that have the same parent codes are called sibling nodes (220 p.74),(209 p.117). An example can be seen in Figure 5.4, as the theme ‘multimedia quality’ is a parent node and all the subheadings (or subthemes) are children nodes. Any two themes within the multimedia are sibling nodes. This process of hierarchy helps in developing a theoretical viewpoint, analysing the data and understanding the participants’ viewpoint, as well as preventing duplication of nodes (220 p.75). After understanding the concept of codes, it can be applied to the data that is being transcribed. This is explained in the next section.

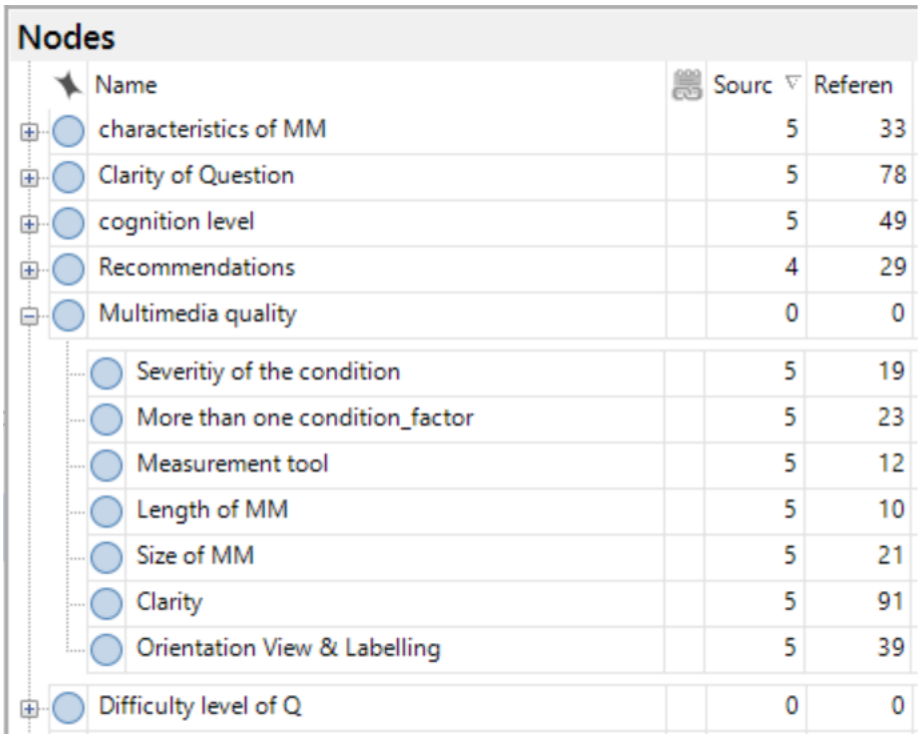


Figure 5.4: Examples of coded themes (parent and child) in NVivo

5.3.4.6.3 The process of transcribing

The discussion of the focus group was carried out in English; however, in all the focus groups, residents would interchange between both languages (Arabic and English) as this is the norm amongst students in the medical field in Saudi Arabia. Translation of the discussion was a complex process that required fluency in both languages, as well as familiarity with its contextual issues.

In total, five focus groups consisting of between four to nine participants (n=33) were conducted in three regions. Each focus group ranged between 1.5-2.5 hours (90 and 150 minutes), generating 11 hours of recordings. All transcriptions were imported into NVivo 11 (QSR International Pty Ltd, 2015), a qualitative software data-management package, which was useful as it helped sort the data and made it easier to create, assign, and merge codes into themes (170, 172, 176). However, the program required training to be able to understand how to use it and get the most out of it. Nonetheless,

the researcher still needed to do the reading, transcribing, and analysis herself. To start transcribing, a non-distractive environment was selected and the use of headphones to muffle outside sounds to hear voices distinctively was used. This allowed the researcher to focus on the audio-recordings and what was being said.

Audio-recordings were transcribed verbatim (word for word) by the researcher, as well as non-verbal utterances (e.g., coughs, laughs) were documented. This provided the researcher with an early sense of the data and an insight into the duration of transcription. The range of ratios that relates tape time to transcription time (hours) varies according to the group discussion and number of participants. Some ratios (tape-time: transcription-time) that have been reported ranged between 1:3, 1:5, 1:6 and 1:10 for the group discussions (175). The researcher's own ratio experience for tape time to transcription time (hours) was 1:180 for the first group discussion, increasing to 1:300 for longer discussions. That meant that every ten minutes of audio-recording required two hours to transcribe. This prolonged duration was because of the bilingual discussion between participants that required frequent pausing, slowing the pace of discussion and repeating the discussion to ensure a complete verbatim transcription was delivered.

When translating, not all concepts can be extracted to another language due to the structure and complexity of the languages that are different (209 p.100). Therefore, not everything can be translated literally (209 p.99). As verbatim translation from Arabic to English will result in ungrammatical English, meaning-based rather than word-for-word interpretation could be used (209 p.100). Each focus group was first transcribed as exactly heard "verbatim transcription" with residents interchanging between Arabic and English to allow for returning to the data at a later stage. This was done using NVivo11 with the contemporary Arabic writing system called "Arabizi",

which is described by Yaghan (2008) as ‘a slang term describing a system of writing Arabic using English characters. This term comes from two words “arabi” (Arabic) and “englizi” (English) and is a text messaging system used over the net and cellular phones (225). In this system, the English letters are used to draw the sounds of the Arabic letters when the same character set was available for both. When the English letters did not have a set for the Arabic letter, it was represented by English numbers that had the same shape of the Arabic letter (225). Figure 5.4 demonstrates examples of these letters taken from the figure presented in Yagha’n (2008 p.43) article (225).

	Arabizi possibilities		
	numerals	letters	
ء	2		
أ	2	a	
إ	2e	e	i
ؤ	2	o	
ا		a	
آ	2a	aa	
ب		b	

Arabic	Arabizi possibilities		
	numerals	letters	
ش		sh	ch
ص	9		
ض	9′	d	
ط	6	t	
ظ	6′	th	
ع	3		
غ	3′		

Figure 5.5: Examples of Arabizi letters from Yaghan (2008) p.43

Arabizi is accepted by most Arabic-speaking people and is growing. This was used because of the difficulty of switching between Arabic writing (direction right to left) and English writing (direction left to right) in NVivo, and because NVivo kept crashing after every other switch. In order to capture the value of what was said, it had to be written in a coherent fluent text before being translated and, therefore, Arabizi was used. In addition, it was less confusing using one set of language keys “English” when listening

to the recording and reading the transcript to follow, as there was a lesser chance of missing something that was said (225).

Because well-constructed sentences and the following of grammatical rules in writing when transcribing is not that straight-forward (220 p.14), most of the dialect words, terms, and grammatical expressions were preserved in the transcription. As this study was mostly concerned with the factual content of what residents were saying rather than with details of language and expression, tidying up the grammatical aspect of the transcript is considered acceptable (220 p.14). In addition, all of the Arabic transcriptions of the residents were highlighted by making it italic in order for the reader to be able to identify which part of the text was originally in the Arabic Language. Only on a few occasions did the researcher have to adjust the grammatical aspect of the sentences after translation in order for it to be understandable and have fluency when one would read between the translated Arabic section of the resident's words and their own words in English. Figure 5.5. demonstrates how the original sentence containing both languages were transformed to English using the Arabizi lettering system and then again translated into English.

Transcription text	Translation method
أنا to me زي ما قالوا الشباب ال video كان ال أكثر شي videos	Arabic writing starts from right to left Translation of both languages as heard
ANA to me Zay MA GALU ALSHABAB IL videos, Kan Ilvideos Akthar Shay)	Translating Arabic to English using Arabizi
For me, it's *like what the guys said*, *the* videos, *the* videos *were the most thins*	Translation to English

Figure 5.6: Example of transcription using Arabizi

The transcription proved to be a lengthy process as illustrated in the literature (170, 217) but was important to gain a closer familiarity with the data and be sure of its accuracy against the original and translated recordings (170, 217). Validation during the transcription phase was performed by continual referral back to the original

transcripts. The transcripts were read whilst listening to the original recording, which was an important step (209, 215 p.79). The audio was listened to on more than one occasion and at separate times from two different recordings to ensure that all conversations were captured. This is important for researchers to do, to be distanced from the analysis process for a period of time so that their perceptions do not influence the themes that are generated (170, 215). This process was difficult to be done by another researcher as it required the familiarity with a medical background, fluency in both languages, as well as familiarity with the different dialects (209 p.100). And as noted earlier that translation to English could not possibly reflect the exact meaning in Arabic therefore, during transcription when needed, a list of common informal Arabic phrases and words were compiled and reviewed with an EM colleague and other peers who were fluent in English and had medical education backgrounds. This was to enhance reliability and to ensure that the meaning of these phrases was appropriately translated. In addition, to ensure that the coding was appropriate after the initial transcription was complete, the researcher went back to the transcripts and made sure that the coding was aligned with the labelled transcripts in the original recordings (220 p.78).

This explained the process of transcribing only highlighted a section of the whole process of thematic analysis. It was important to first understand the concept of thematic analysis and to understand what a code was and how to apply it to text. However, before one is able to apply the codes to it appropriately, the process of transcribing and translating the data first and how to go about it needed to be explained. The following section describes how thematic analysis was undertaken using a thematic analysis guide.

5.3.4.6.4 Applying thematic analysis

Thematic analysis was undertaken in this research using a six-phase guide described by Braun and Clarke (2006) (217). These stages are also mentioned in other articles some as guidelines and others as headings (175, 215, 217, 218, 220). However, Braun and Clarke's guide was the most systematic and their six phases of thematic analysis are: 1) Familiarizing with the data; 2) Generating initial codes; 3) Searching for themes; 4) Reviewing themes; 5) Defining and naming themes; and 6) Producing the report. These stages are presented in more detail in Table 5.10 with the researcher's approach to analysing the focus group data for each stage.

Table 5.10 Focus group analysis using thematic analysis (Braun & Clarke, 2006) (217)

Thematic Analysis Phase	Meaning and Explanation
Phase 1: Familiarizing with data	Transcribing data (if necessary), reading and rereading the data and noting down initial ideas
The researcher immersed herself in the data to be familiar with its content through writing the transcription, reading and re-reading the data, making marginal remarks (ideas and understanding about the data), and searching for meaning and patterns. Notes were taken about important key points, instructions to oneself to seek further clarification about an idea within the data, links to theories in the literature, and references to data in other transcripts or within the same transcript. The researcher also formed ideas for initial coding in this phase to help in the next phase of coding.	
Phase 2: Generating initial codes	Coding interesting features of the data in a systematic fashion across the entire data set, collating data relevant to each code
This starts after reading and familiarising yourself with the data and having generated a list of initial ideas about what is interesting. In this phase, the researcher started producing initial codes from the data by searching the data for interesting aspects and repeated patterns that were common across residents and items. The process of coding in its initial phase was more of a descriptive code where labels were added to what the participants described. Notes and annotations were also used during the process of coding as a way of theorising and commenting on the general development of the conversation and analysing thoughts about the codes to make them clearer. A long list of initial codes was identified and formulated in this stage.	
Phase 3: Searching for themes	Collating codes into potential themes, gathering all data relevant to each potential theme
Themes are developed in this stage where interpretive analysis of the data occurs. The researcher looked at the coded data again to consider for patterns and relationships within a transcript and across transcripts. Further review of the initial coding was through categorising the text into related themes and for underlying meaning. All relevant codes and data extracts were sorted into potential themes for a broader level of analysis. Here, the researcher analysed how different codes are combined to form an overarching theme, and what relationships lied between codes (parent and child) and themes in order to help develop subthemes for phases 4 and 5.	

Phase 4: Reviewing themes	Checking if the themes work in relation to the coded extracts (Level 1) and the entire data set (Level 2), generating a thematic map of the analysis

Here, reviewing and refinement of the themes occurred. Reviewing themes involved reading all collated extracts for each theme to decide whether or not they form a coherent pattern. After reviewing the themes, they were reviewed again in relation to the entire data set and were refined (e.g., coded any additional data that were missed, collapsed related themes into one, or separated a theme into two). Additional codes were identified during the process of coding the data by questions (e.g., MM-TXT question one across the five focus groups to track the whole picture of the item by all the groups), as well as by type of questions (multimedia or text). Finding additional codes at this stage was expected as coding is an ongoing iterative process (215, 217, 222). A thematic hierarchy demonstrates the relationship between codes (217). This was created in Nvivo with relations of themes and subthemes and presented as a table in the results section in Chapter 6.

Phase 5: Defining and naming themes	Ongoing analysis for refining the specifics of each theme and the overall story that the analysis tells, generating clear definitions and names for each theme

After having an overall theme (table or map), the researcher then defines the themes in the sense of what it captures and how it fits in with the overall data and research question. This is done by going back and checking with the extracted data and identifying which themes contained sub-themes. That means that codes were linked and unified in one general theme with related subthemes under it. Deciding on the names of the themes for the final analysis was considered here. Further, into the analysis, the frequency of codes and themes were calculated (i.e., quantitizing the data). The code frequency aids researchers in having an objective measure of the prevalence of topics or relationships between and within groups but does not indicate its importance (171, 176).

Phase 6: Producing the report	The final opportunity for analysis, selection of vivid, compelling extract examples, final analysis of selected extracts, relating back of the analysis to the research question and literature, producing a report of the analysis.

After having the final themes, the final part of the focus group process is selecting relevant and appropriate quotes to be presented for each theme (results section) and reporting the findings in the discussion section, where the researcher reflected on the results in the light of empirical and published work. Evidence was provided using data extracts and quotes to support each theme. The extracts illustrate the analytic points about the data (217). Descriptions and arguments related to the research question and with results from other methods were also provided (this is found in the results and discussion section).

Although the stages carried out seem to be simple, in reality, analysing the data was an iterative, rigorous, lengthy and sometimes frustrating process that involved listening to the recordings several times, identifying different voices, transcribing, translating, re-reading the transcripts a number of times, categorising the data and making sense of it, coding the statements, developing appropriate themes and selecting appropriate quotations for the write-up phase. The following section is related to the sixth and seventh methods listed in Table 1.1 and used in this research and is related to validity framework and legitimation.

5.3.5 Validity framework and legitimation

Sailing through the literature on validity that was described earlier in Chapter 4, it can be noticed that there are various conceptual frameworks and major thinkers in the area of validity and that there has been a noteworthy change in how experts in measurements and psychometry conceptualize validity (37). In the past century, key players in the field such as Messick and Kane gave detailed reviews about validation (157). This research is methodologically based on the literature that explores the difficult concepts of assessment validity and draws on the American Psychological Association, as the overall structuring framework described by Downing (164), and noted by the Standards for Educational and Psychological Testing (explained in Chapter 4) (155), to describe and report the research while critically drawing on a new 'operationalising' framework offered recently by the Cambridge Assessment Group (35). In order to measure the effectiveness of multimedia MCQ items in this setting, the validity of this procedure should be undertaken. Therefore, a mixed-methods approach to the Cambridge Validity Framework was used to draw on the strengths of both qualitative

and quantitative data collection. This also aligns with the research design of mixed methods taken up in this research. An understanding of the consequences of a new assessment intervention “multimedia in MCQs” within the newly published validity framework from the Cambridge Assessment Group was evaluated and explored.

In addition, for the purpose of this research, there were two types of validity checks that were applied. The first validity check was using the Cambridge Validity Framework to conduct the test and validate its results. This framework supposedly covers important points in each step of the testing process to take into consideration when conducting the examination as demonstrated in Table 5.5. The second validity check was for the mixed-methods research approach called “legitimation”; which evaluates the whole research process from data collection to presentation of results. Both of these validity checks are considered part of the research methods that were listed in Table 1.1.

5.3.5.1 The Cambridge framework

This research applies the Cambridge Validity Framework, around the use of multimedia in MCQ written examination. This framework proposes steps and associated requirements for validity evidence for assessment holders to use in developing their own assessment needs. The Cambridge Assessment Group, responsible for delivering the Cambridge Board’s A level examinations internationally, has proposed an operationalised approach (i.e., identifying appropriate instruments to be used to measure the construct) to the earlier work of Kane (2006). The framework is structured to be accessible to all those who are involved in the development of assessment and facilitates its application by suggesting a set of possible methods and sources to be used as evidence for the validity of assessment at all levels of the process. The Cambridge

Assessment Group's framework proposed an operationalised approach based on Kane's latest argumentative-based approach, which consists of a number of research questions, each linking to a chain of inferences. To be able to understand their framework, Figure 5.7 depicts the Cambridge framework developed in their article and was adopted and used for the purpose of this research (35 p167), and Appendix 18 presents the Cambridge framework with the proposed methods as described in their article.

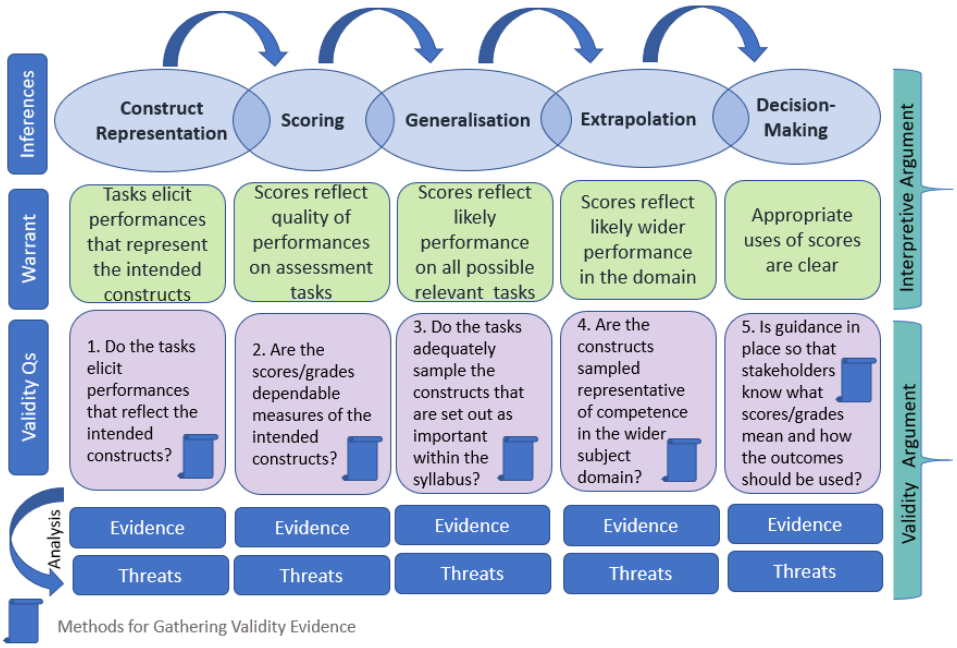
Interpretive Argument		Validity Argument		
Inference	**Warrant justifying the inference**	**Validation question**	**Evidence for validity**	**Threats to validity**
Construct representation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?		
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/grades dependable measures of the intended constructs?		
Generalisation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?		
Extrapolation	**Scores/grades reflect likely wider performance in the domain**	4. Are the constructs sampled representative of competence in the wider subject domain?		
Decision-making	**Appropriate uses of scores/grades are clear**	5. Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?		

Shaw S, Crisp V, Johnson N. A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assessment in Education: Principles, Policy & Practice. 2012;19(2):159-76.

Figure 5.7: Cambridge framework for the argument of assessment validation

As seen in the figure, the Cambridge framework demonstrates the argument-based approach and, like Kane's framework, it consists of two parts: a) an Interpretive argument and b) a validity argument. The interpretive argument itself also contains two parts

inferences and warrants. Inferences or claims are ‘statements of claimed inferences from assessment outcomes and the warrants which justify the inferences’ as described by Shaw, Crips, and Johnson (35). The five inferences move in a chain of inferences from one to the other, from construct to decision making, as illustrated in Figure 5.8.



Cambridge Validity Framework :Shaw S, Crisp V, et al. (2012). "A framework for evidencing assessment validity in large-scale, high-stakes international examinations." Assessment in Education: Principles, Policy & Practice 19(2): 159-176.

Figure 5.8: Illustration of the Cambridge framework

The first step is transforming the task to test performance (construct representation); second, from test performance to test scores (scoring); third, moving from the test-takers’ observed score to their universe score or test competence (generalisation); fourth, (extrapolation) is moving from the test competence to the domain competence (in the real-world setting); and last is (decision making) from domain competence to trait competence. All inferences are represented in Kane’s framework except construct representation. For each inference, there is a warrant (which is used to justify that the inference is true). The warrant in this framework is statements that need to be backed up

by evidence in order to be able to pass the validity argument. Evidence collected should include strength and justification to the inferences that were made and evidence for weakness through identifying possible threats and areas for improvement. Appendix 18 gives examples of methods to aid in collecting these evidences. The second part of this framework is the validity argument, which evaluates and analyses all the gathered evidence regarding these inferences. It contains research questions to help structure the argument while still linking it to the appropriate inference. The evidence collected would either support validity or suggest potential threats to it (35).

It should be clearly noted that although all of the inferences may be valuable, not all forms of validity evidence are needed for an assessment procedure to be an acceptable measure of performance and that it is never possible to explore all areas of validity evidence. Generally, programs of research are established to explore all areas over time (37, 157). However, in summative assessment where the stakes are higher (e.g., licensing and board examinations), more rigorous validity evidence is necessary to be gathered in order to defend the decision made on the examinee's performance (37). In this research, only the first three inferences were applicable, as extrapolation and decision making require further follow-up and longitudinal studies. For example, extrapolation requires studies establishing relationships between the results and other variables. Regarding decision making, evidence in this area can be gathered from monitoring the outcomes of decisions (success or error) and evaluating the intended and unintended consequences through long-term follow-up studies, qualitative studies and considering the impact on learners, raters, and others (37).

Simply put, the framework involves questions to be addressed in every step of the test

development from training item writers, constructing the blueprint, developing exam items, to selecting appropriate methods, delivering the test, and analysing and interpreting the results. This is to ensure that inferences made from results are valid.

5.3.5.2 Framework application

Theorists are continuously thinking and re-thinking on what the word validity means and to what it applies to (for e.g., an item, a test, a score, a decision, a consequence, a policy an inference, etc.) while practitioners won't. This distance between the theoretical aspect of validity and what is practised in reality is not fully paved (158). Therefore, this research draws on one of the published frameworks to validate the assessment process and to report on its use, applicability and feasibility for others who wish to know how it is done.

As explained, in this research, the first three steps were applied to the test development process. The first inference ‘construct representation’, is concerned with how well the construct under study is represented from the computerized MM-TXT examination and its results. Methods of collecting validity evidence are from experts and examiner’s classification of items and their relevancy to residents’ training and curriculum. Another source of validity evidence is from the residents’ results, item analysis parameters from their test scores, and evidence of the presence of CIV. This was reflected in Figure 5.1 as previously illustrated, where evidence for construct representation could be seen to be gathered at multiple stages of the test development process and the research phases (quantitative and qualitative phases). This can be better viewed from Table 5.11, where the inferences are laid out against the research phases and the sources of evidence gathered from implemented methods in this research.

The second inference ‘scoring’ reflects the quality and reliability of the test performances and results from all phases of the research project, as can be seen in Table 5.11. The third inference, ‘generalization’, differentiates performance or competence in the ‘test worlds’, also known as ‘universe of assessment’ from performance in the ‘real world’, which is extrapolation. The logic of the universe of assessment is that in the test world, there is an infinite number of possibilities of items that could be selected to assess the domain under study. The test items that are eventually chosen for a given test represent only a sample of the items from this universe of possibilities (157). Therefore, generalization is concerned with theoretically knowing how likely would the selected sample items represent all possible items in the relevant universe of assessment, and how likely performances in these exams would reflect on other tasks that were relevant to the construct (i.e., generalizing takes us from the test sample to the entire assessment universe, while extrapolation takes us from the assessment universe, performance to the real-world performance (domain competence) (157).

Table 5.11 Framework inferences, research phases and sources of validity evidence

Inferences **Research Phases**	**Sources of Validity evidence**
Construct representation	
Exam development (phase I) Exam development (phase II)	BP review, item classification by experts, item cognitive level, CBT guidelines and issues
Data analysis (phase I) Data analysis (phase II)	Quantitative results: Residents' scores, statistical analysis on item level and cognition, item analysis and sources of CIV.
Data analysis (phase III)	Qualitative results: Questionnaire and focus group discussions results
Scoring Data analysis (phase I) Data analysis (phase II)	Quantitative results: Review scoring and marking procedures, standard settings (SS), review of item classification, level and cognitive level by expert, reliability and statistical analysis on exam components and residents' levels.
Data analysis (phase III)	Qualitative results: Feedback from residents' input on survey and focus group discussions results
Generalization	
Exam development (phase I) Exam development (phase II)	Quantitative: Appropriate BP and test domain sampling, statistical analysis of cognitive level and expert review, generalizability study Qualitative: Empirical studies with similar results about reliability and reproducibility, focus group discussion results

Evidence of extrapolation is given in Appendix 18 and primarily come from further interviews with experts, observations of actual skills and tasks performed in the field, and think-aloud studies given by experts while performing the tasks (157). It is important to remember that because validity is contextual, it may not be transferable and can only be transferable to another assessment process or instrument if it is kept close to its original implications (content domain, learners context, performance context of the assessment

data, etc.) (37). The closer one contextual situation is to another, the more evidence one has on validity to apply from one context to another (37).

The last inference ‘decision making’, moves us from the real-world performance to an interpretation about that performance, how it’s used for decision making, its usefulness and impact on the learner, stakeholder and society (i.e., consequences). This was not applicable to this research and until now evidence of consequences in the literature is absent (157).

5.3.5.3 Research validity (legitimation) in mixed-methods research

Research validity in mixed-methods research is called legitimation and relates to how true and correct are the inferences that are made from the research results and the research process (methods, methodology, data collection and interpretation) (159). There are various types of validities for quantitative and qualitative methods and they all need to be reviewed and checked for their appropriateness in mixed research. Appendix 19 provides a checklist that was adopted from Johnson and Christensen’s Handbook: Educational Research: Quantitative, Qualitative, and Mixed Approaches (2016) (159 p.104), and outlines multiple questions related to the introduction, methods, results and discussion phases of the research project. This can be used to evaluate the quality of a mixed research study.

To validate the inferences that are made from the study results, validity checks for quantitative, qualitative, and mixed methods were checked. Onwuegbuzie and Johnson proposed eleven types of mixed research validity to be reviewed and involve issues related to the design of the research, methods selected and structure of the whole

research process. These were addressed by applying them throughout the whole research process from data collection, to method and data analysis and can be seen through the guidelines and frameworks that were applied. These types of legitimations are viewed as types of validity for mixed research and reflect the validity framework for the mixed-methods research to present evidence that either strengthens the results or weakens them. In mixed-methods research, conclusions are based on the quantitative and qualitative components of the undertaken research. These inferences should be integrated into a larger meta-inference (159). The eleven types of validity are inside-outside legitimation**,** paradigmatic/philosophical legitimation**,** commensurability approximation legitimation**,** weakness minimization legitimation, sequential legitimation, conversion legitimation, sample integration legitimation, socio-political legitimation, multiple validities legitimation, pragmatic legitimation, and integration legitimation**.** Appendix 20 presents these types of validity checks for mixed-methods research, with their meanings, examples, and their applications in this research (226 p.306-9).

5.4 Conclusion

This chapter demonstrated the research design, methods, and rationale used to collect the appropriate data for this research, as well as the analysis methods taken. The research design was an exploratory mixed-methods design with a sequential quantitative-qualitative timing. The mixed-methods was the most appropriate approach for gathering information from multiple sources (quantitative and qualitative) to have more than one perspective on the data. The previous methods described were used to investigate the effect of multimedia in MCQs to measure higher cognitive levels (i.e., the underlying construct), as a means of increasing the validity of examination. The concept of validity

and frameworks were explained to emphasise the importance of evaluating the methods used to gather evidence, the examination process, as well as the whole research. Various frameworks, checklists, and guidelines were undertaken in each research method to ensure the soundness of their results. The following chapter demonstrates the results of these methods.

Chapter 6: Results

6.1 Introduction

This section relates to the results of the MCQ examination formats that were used during the pilot and test study and justification for their combination. Two high-stakes examination tests were formed and conducted in the years 2013 and 2015 with emergency medicine (EM) residents. The tests consisted of 100 marked questions (residents' promotion exam) in which all residents would need to pass in order to get promoted to their next training year. In addition to the marked questions, (unmarked) multimedia text-matched questions were added, 30 questions in the 2013 EM exam and 50 questions in the 2015 EM exam. All residents took the same 100 marked questions, but they differed in the unmarked ones receiving either multimedia questions (MM form) or the text-matched questions (TXT forms).

6.1.1 Justification for data combination (data manipulation)

A t-test was conducted to demonstrate if there were any differences between the two residents' groups of different years (2013 and 2015) who have taken the 100 promotion questions. This allowed us to see if data from both groups were similar, and hence, would allow for the combination and merging of data sets to have a larger sample size for analysis. Both examination setups had the same methods for the preparation and execution of the multimedia and text formats. Outliers were tested for, and were removed from the population and results were retested again. All testing yielded that there was no significant difference when outliers were removed, and the means were not altered. In addition, with the combination of data for both years and a sample population of more than 30 participants, the effect of outliers was minimal (227). In this study, all subjects

were kept after demonstrating that their removal had no effect on the data. The following section is for data manipulation that demonstrates a series of t-tests and outlier testing for the items on each year separately, to justify their combination. The final results are reported from the combined data together.

6.1.1.1 Series of t-tests for data combination

A series of independent and paired t-tests were conducted on the total test scorers of the promotion items and the pilot (beta) items to explore if there were any differences among the data by groups, forms, years and exams. These are all demonstrated in Table 6.1. An independent sample t-test was conducted to compare the total test scores on the marked promotion items between residents from both years (2013 and 2015). No significant difference between the groups was found, (t = .116, df = 162, p =0.91) for the mean scores of 2013 (70.8$\pm$8.6) and 2015 (70.6$\pm$9.2), as shown in Table 6.1.

Student's t-tests were then undertaken to see if there were any differences in score levels on the same marked promotion items between groups by forms (either text or multimedia). The first groups were those who had been assigned to take the text version of the examination in the year 2013 and 2015 while the second groups were those who had been assigned the multimedia questions from both years. The mean test scores for the text group form was 70.2$\pm$8.4 and 71.2$\pm$9.3 for the multimedia group with no significant differences found between them (t = -.710, df = 162, p =0.48).

Regarding the pilot (beta) items, an independent sample t-test was also carried out comparing scores of 80 TXT-MM items for residents by years (2013-2015) and there were no differences in scores between the TXT-MM groups of the year 2013 and 2015. To see

if there were any differences among a single type of form, a paired t-test was carried out to see if there were any differences between the text group of 2013 and the text group of 2015. A p-value of .572 showed no differences. Similar results were yielded for the multimedia groups of 2013 and 2015 (see Table 6.1). Therefore, text items of both years were combined for the promotion questions, as well as for the unmarked multimedia and text items. An independent sample t-test between the MM group (all residents of 2013+15) and the TXT group (all residents) was done with 83 residents taking the pilot text format and 81 taking the pilot multimedia format. Mean test score for the text group was slightly higher (X=75.36 $\pm$ 8.26) than the multimedia group (X= 73.45 $\pm$ 9.11) but no significant differences were detected between both groups (t = 1.410, df = 162, p =0.16).

Table 6.1: Series of independent and paired sample t-tests on promotion and pilot exams

Resident Characteristics			**Item Characteristics**		**Mean (SD)**	**t (df)**	**p-value (2-tailed)**
Group	**Year**	**N**	**Type**	**N**			
TXT + MM	2013	80	Marked	100	70.8±8.6	. 116 (162)	0.91
TXT + MM	2015	84	Marked	100	70.6±9.2		
TXT	2013+2015	83	Marked	100	70.2±8.4	-.710 (162)	0.48
MM	2013+2015	81	Marked	100	71.2±9.3		
TXT + MM	2013	80	Pilot	80	73.83 ± 8.31	-.840 (162)	.402
TXT + MM	2015	84	Pilot	80	74.98 ± 9.10		
TXT	2013	40	Pilot	30	74.83 ± 7.28	--	.572
TXT	2015	43	Pilot	50	75.86 ± 9.14		
MM	2013	40	Pilot	30	72.84 ± 9.21	--	.552
MM	2015	41	Pilot	50	74.05 ± 9.08		
TXT	2013+2015	83	Pilot	80	75.36 ± 8.26	1.410 (162)	0.16
MM	2013+2015	81	Pilot	80	73.45 ± 9.11		

6.1.1.2 Descriptive statistics for promotion items

Descriptive statistics for the groups of text and multimedia are presented in Table 6.2 below for the marked promotion items. Descriptive statistics for the unmarked pilot items are given later. The text group comprised of 83 residents (40 from 2013 and 43 from 2015). The number of residents who received the multimedia form was 81 (40 residents form 2013 and 41 from 2015). The mean test score for the text group was 70.2±8.4 and for the multimedia group 71.2±9.3.

Table 6.2: Descriptive statistics for MM and TXT groups on the promotion items (2013, 2015)

Variables	Text Group on 100 promotion items (n=83) *	Multimedia Group on 100 promotion items (n=81) *
Mean ± SD	70.2±8.4	71.2±9.3
Median	70.71	73
Skewness	-.557	-1.873
Kurtosis	-.526	4.108

*n represents the number of residents

6.1.1.2.1 Normality test

Some of the data was found to be skewed and was tested for normality using the Kolmogorov-Smirnov test (see Table 6.3). It was found to be not normally distributed. Further tests were applied using the Mann-Whitney U test (test of the median) as it is more stable than the mean. The test was non-significant, as shown in Table 6.4 (p = 0.32), which indicated that the groups are the same and that the results were similar to the test of means and, therefore, the independent sample t-test was used and presented. In addition, using a sample size that is more than 30 will give a normal distribution, and

the test of the mean and median will give similar results (227). Therefore, the data were treated as being normally distributed.

Table 6.3: Test of normality (Kolmogorov-Smirnov) for the promotion items

	Group	Statistic	df	Sig.
Score	1 TXT	.109	83	.016
	2 MM	.178	81	.000

Table 6.4: Mann-Whitney U test for the promotion items

			Text	Multimedia
N			83	81
Percentiles	**(IQ)**	**25**	64.0	68.3
	Median	**50**	70.7	73.0
	(IQ)	**75**	77.0	76.8
P value (Mann-Whitney U) *			0.32	

*Significant level at 0.05

The illustrated Q-Q plot of scores for both groups (multimedia and text) on the promotion exams are presented in Figures 6.1 and 6.2, respectively, and demonstrate that deviations from the straight line are minimal. Even with the removal of the deviated data the scores were unchanged, which indicates that the data can be treated as a normal distribution.

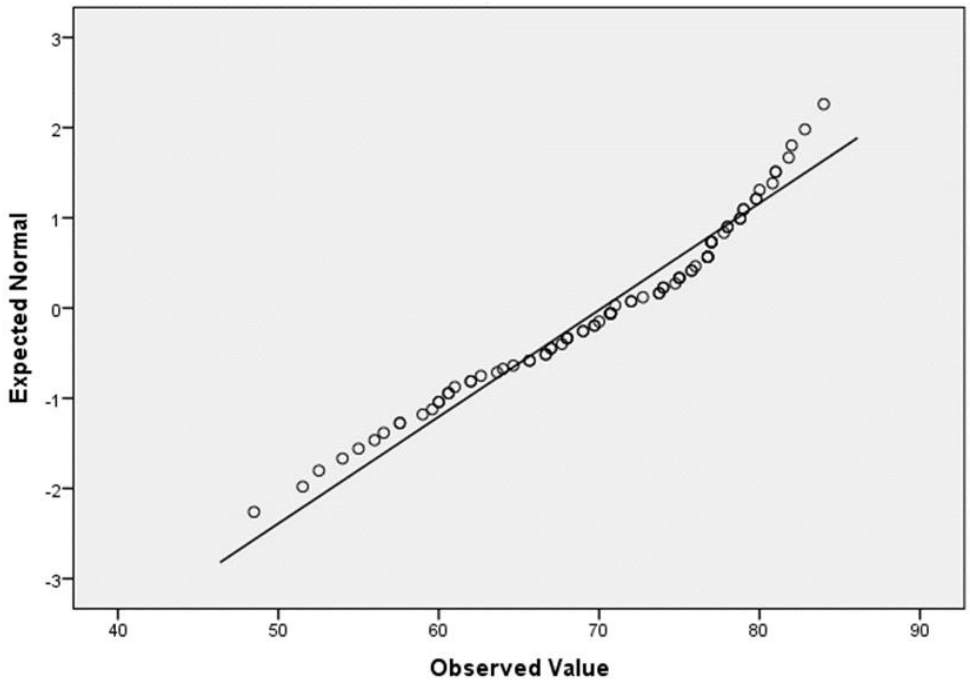


Figure 6.1: Normal Q-Q plot of scores for the text group taking promotion items

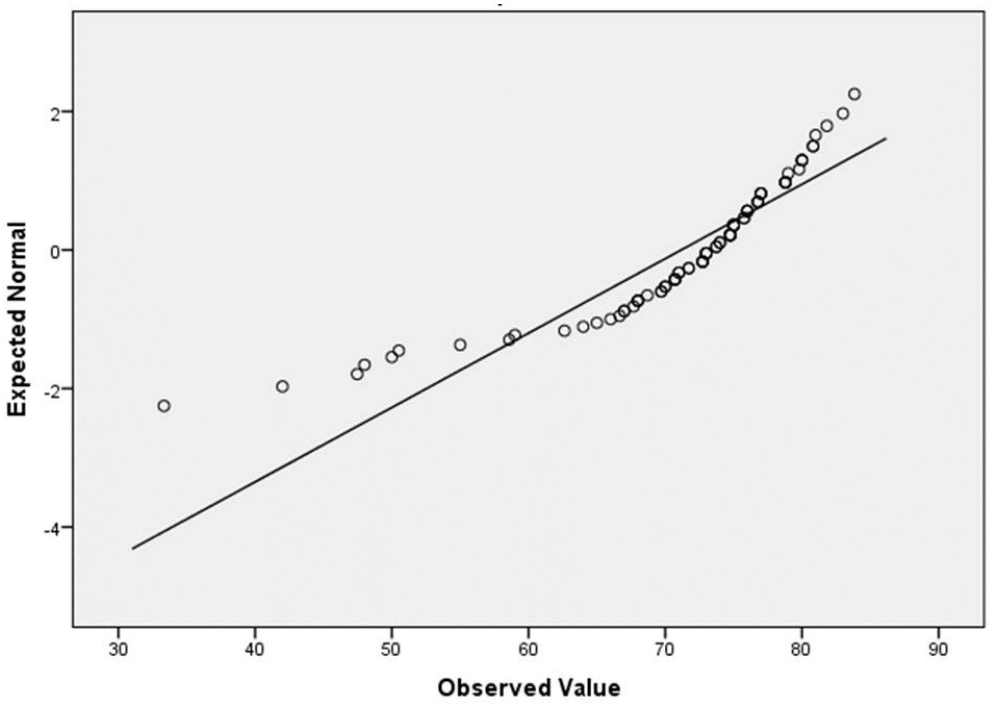


Figure 6.2: Normal Q-Q plot of scores for the multimedia group taking promotion items

6.1.1.2.2 Outlier testing

The data were examined for the presence of outliers among residents' performances on the promotion exam and were identified and removed, see Table 6.5 and Figure 6.3. The t-test was re-calculated and results showed that after their removal, there was no difference between mean test scores of the two groups (p=0.15). Therefore, it was decided that they would be kept, as the presence of the whole data set would give a more realistic picture of the exam.

Table 6.5: Mean scores of original data and after outliers removed on the promotion exam

	Original Data	**After Outliers removed**
Text	70.2 (83)	70.2 (83)
Multimedia	71.2 (81)	72.0 (79)
P - value	0.48	0.15

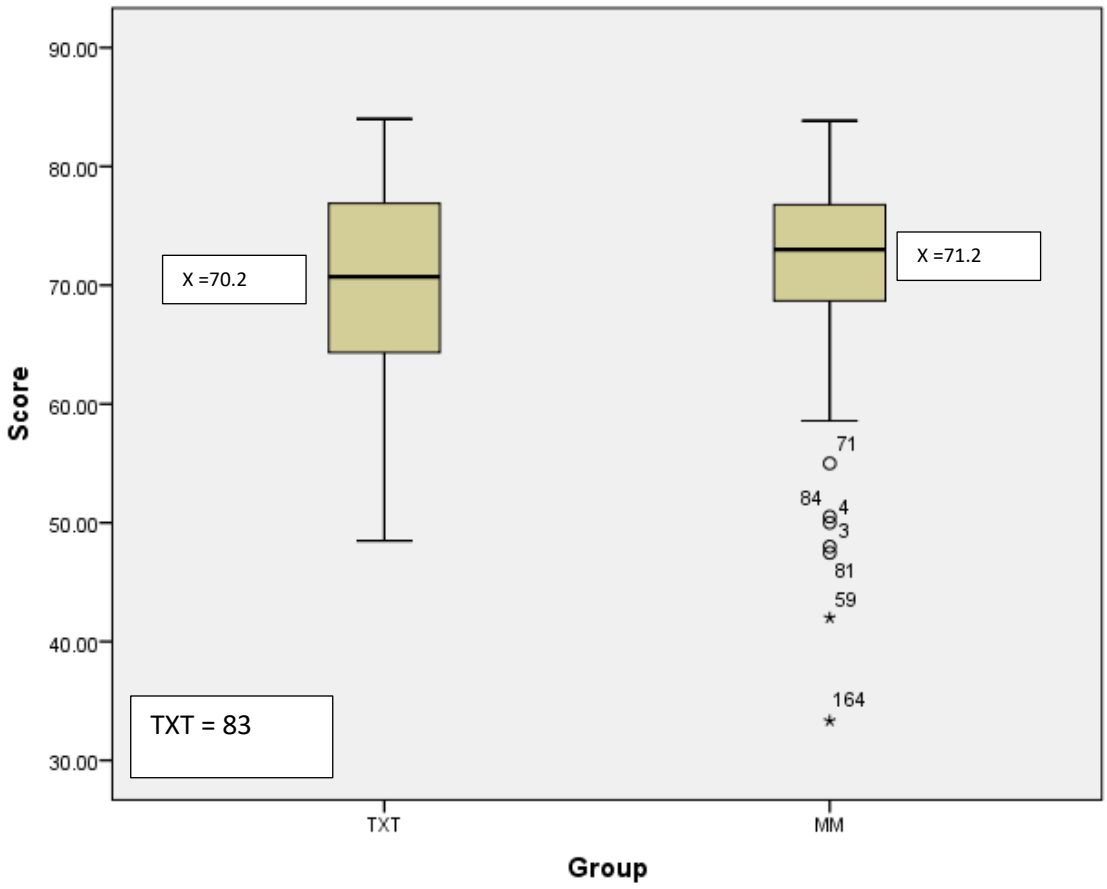


Figure 6.3: Box plot demonstrating the presence of outliers in both groups taking the promotion items

6.1.1.3 Descriptive statistics for pilot exam

The following data are the same series of tests conducted (descriptive, normality, and outlier testing) but on the unmarked, pilot (beta) questions.

6.1.1.3.1 Normality test

Results from Table 6.6 demonstrate a slight skewness in the data, and, therefore, the data needed to be tested for normality using the Kolmogorov-Smirnov test.

Table 6.6: Descriptive statistics for MM and TXT groups on the pilot items (2013, 2015)

Variables	Text Group on 80 pilot items (n=83) *	Multimedia Group on 80 pilot items (n=83) *
Mean ± SD	75.36 ± 8.26	73.45 ± 9.11
Median	76	74
Skewness	-0.375	-1.086
Kurtosis	.151	2.185

*n represents the number of residents

Results from Table 6.7 were significant for both groups, which indicated that they were not normally distributed. Therefore, a non-parametric test was conducted as done before using the Mann-Whitney U test (see Table 6.8). The test was non-significant (p = 0.22), and the medians of the scores were the same across categories of forms, which indicated that the groups are the same and that the results were similar to the test of the mean. Therefore, the independent sample t-test was also used and presented as the main result.

Table 6.7: Test of normality (Kolmogorov-Smirnov) for the pilot items

	Group	Statistic	df	Sig.
Score	1 TXT	.122	83	.004
	2 MM	.118	81	.007

Table 6.8: Mann-Whitney U test for the pilot items

			Text	Multimedia
N			83	81
Percentiles	**(IQ)**	**25**	68.0	70.0
	Median	**50**	76.0	74.0
	(IQ)	**75**	82.0	80.0
P value (Mann-Whitney U) *			0.216	

*Significant level at 0.05

Again, the Q-Q plot of scores for both groups (MM and TXT) on the unmarked pilot items are shown in figures 6.4 and 6.5 and demonstrate data clustering alongside the line. Exploration and removal of the deviated data were done with no difference found. Since the deviations from the straight line were minimal, having a sample size that was more than 30, and after conducting the non-parametric test that was not significant, the data were treated as normal distribution and parametric tests were used.

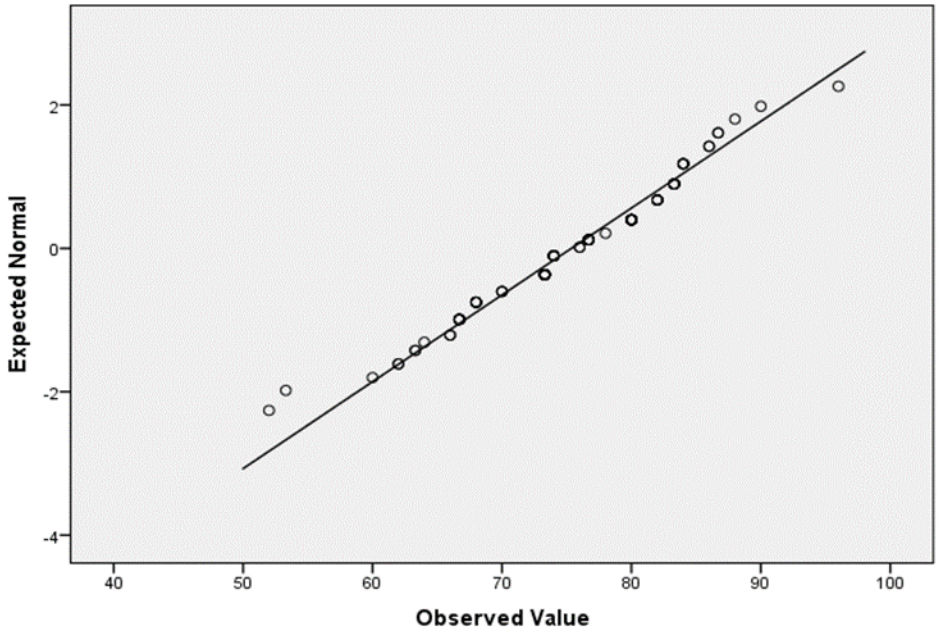


Figure 6.4: Normal Q-Q plot of scores for the text group taking pilot items

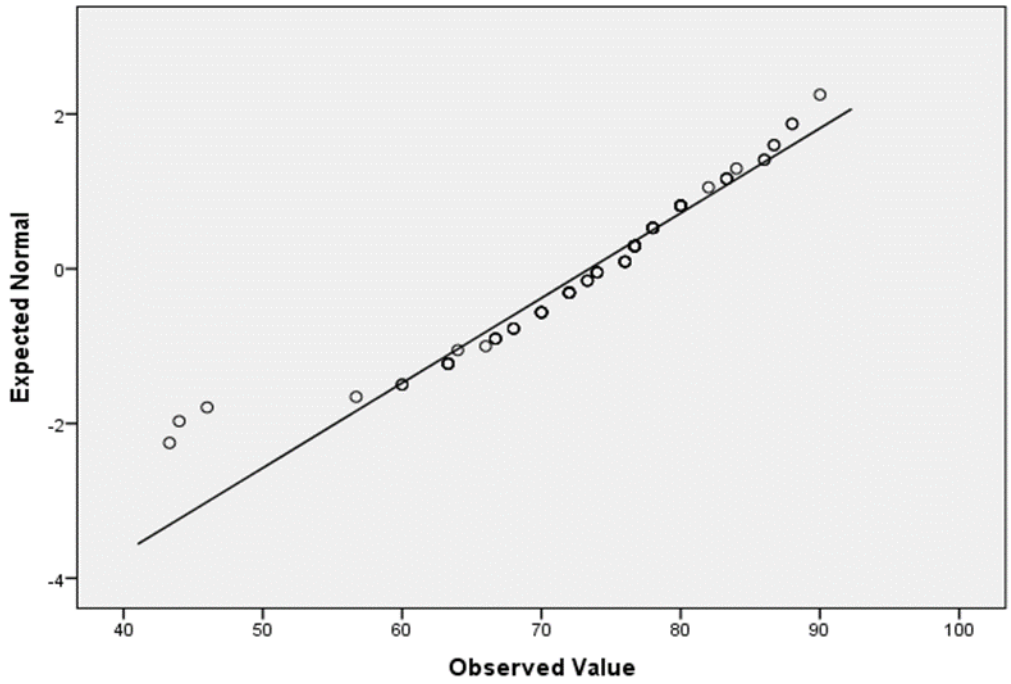


Figure 6.5: Normal Q-Q plot of scores for the multimedia group taking pilot items

6.1.1.3.2 Outlier test

Similarly, the unmarked pilot test items were examined for the presence of outliers (see Table 6.9 and Figure 6.6). No significant outliers were noted in the MM-TXT group. Even when trying to remove the three outliers presented in Figure 6.6 and re-calculating using the independent sample t-test, no changes were found and results showed that after their removal, there were no differences between mean test scores of the two groups (p=0.52). Therefore, it was also decided to keep these individuals, as the presence of the whole data set would give a more realistic picture of the exam.

Table 6.9: Mean scores of original data and after outliers removed on the pilot exam

	Original Data	**After Outliers removed**
Text	75.36 (83)	75.36 (83)
Multimedia	73.45 (81)	74. 57 (78)
P - value	0.16	0.52

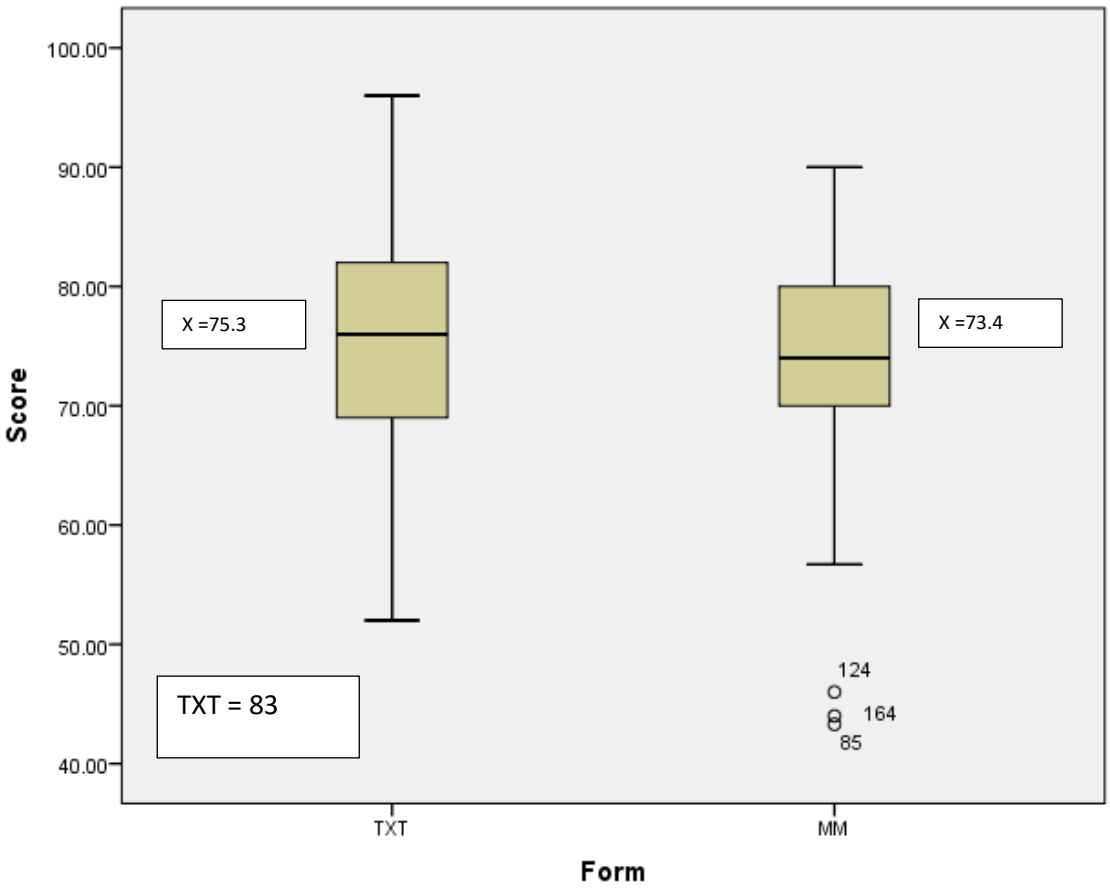


Figure 6.6: Box plot demonstrating the presence of outliers in both groups taking the pilot items

6.1.2 Combined Data Results of the Residents and Items

The following results presented will be focused on the combined data of pilot (beta/unmarked) questions that were being tested (80 multimedia and text-matched questions), which is the focus of this research study, as well as the combined residents' scores on these items.

6.1.2.1 Combined Data Results of residents' examination total scores

A total of 164 emergency medicine residents participated in this research study and took the Promotion-MM-TXT matched questions examination. Eighty residents in 2013 (40

taking the text form and 40 the multimedia form) and 84 residents in 2015 (43 receiving the text form and 41 the multimedia form), see Table 6.10.

Table 6.10: Residents and items count for 2013 and 2015

Year of test	Text Form Exam			Multimedia Form Exam			Total Residents
	Residents	Promotion Items[1]	Pilot[2] Items	Residents	Promotion Items[1]	Pilot[2] Items	
2013	**40**	100	30	**40**	100	30	**80**
2015	**43**	100	50	**41**	100	50	**84**
2013 + 2015	**83**	200	80	**81**	200	80	**164**

[1] This exam had the marked items included in their final score
[2] These items were unmarked not included in their final score

Participating residents were from different training levels, comprising of both juniors and seniors (R1, R2 and R3). The 4th year residents in their final training year (R4) were not included in the study in accordance with new SCFHS examination rules and regulations that exempted them from entering the final promotion examination and having only to take their final board examination. As seen in Table 6.11, 83 of these residents received the text format of MCQs and 81 received the multimedia format; 77% of residents who participated in the examination were male (n=127) and 23% were female (n=37). The bulk of residents were from the Central region (68%), with the remaining residents equally distributed from the Western and Eastern regions (16%). Out of these 164 residents, 62 were in their first year of residency training (R1), 41 were in their second year, and 61 were in their third year (Table 6.15).

Table 6.11: Mean test scores of pilot (unmarked) MM-TXT questions organised by forms

Test score	N	Minimum	Maximum	Mean (SD)	t-test (p-value)
Text	83	52	96	75.362 (8.26)	
Multimedia	81	43	90	73.449 (9.10)	0.16
Total	164	43	96	74.42 (8.71)	

Taking a look at the results from the tables, the mean test scores for the text group were slightly higher than those for the multimedia group. However, they were not statistically significant when conducting an independent sample t-test. When comparing mean test scores by gender (Table 6.12), there were no differences in mean test scores between males and females (p =0.302), even by splitting the data into text and multimedia groups respectively (p=.124, p = .99). Results demonstrated weak to no association between gender and mean test scores (Eta = 0.081, Eta Sq = (0.01)).

Table 6.12: Mean test scores of pilot (unmarked) items organised by gender

Gender	All Groups		Text Group		Multimedia Group	
	N (%)	Mean (SD)	N	Mean (SD)	N	Mean (SD)
Male	127 (77.4)	74.797 (9.02)	64	76.12 (8.42)	63	73.45 (9.47)
Female	37 (22.6)	73.113 (7.53)	19	72.80 (7.33)	18	73.44 (7.94)
Total	164	74.42 (8.71)	83	75.36 (8.26)	81	73.45 (9.11)
ANOVA (P-Value)	-	0.302		.124		.998
Eta (Eta Squared)		.081 (0.01)		.170 (.029)		.000 (.000)

Comparing the groups by residency levels, the mean test scores were statistically significant (p < 0.000) among residency levels, see Table 6.13. An analysis of variance was conducted and the effect of residency level was significant, (F (2,27) = 5.94, p < 0.000). The Tukey HSD procedure (table 6.14) revealed that pairwise differences among means of R1,2 and R1,3 were significant, (p < 0 .00). There was a moderate association between mean test scores and residency levels (Eta = 0.406, Eta Sq = (0.16)) with 16% of the variations in the scores explained by variations in the residency levels.

Table 6.13: Mean test scores of pilot (unmarked) items organised by Level

	All Groups		Text Group		Multimedia Group	
Level	**N (%)**	**Mean (SD)**	**N**	**Mean (SD)**	**N**	**Mean (SD)**
R1	62 (37.8)	69.956 (9.35)	34	71.53 (8.17)	28	68.05 (10.46)
R2	41 (25.0)	76.253 (7.77)	21	75.45 (7.35)	20	77.09 (8.30)
R3	61 (37.2)	77.718 (6.56)	28	79.95 (6.73)	33	75.82 (5.87)
Total	164	74.42 (8.71)	83	75.362 (8.26)	81	73.449 (9.10)
ANOVA (P-Value)	-	0.000		.000		.000
Eta (Eta Squared)		.406 (0.16)		.442 (.195)		.437 (.191)

Table 6.14: Tukey HSD (post Hoc) multiple comparisons between residency levels

level	N	1	2
R1	62	69.9565[b]	
R2	41		76.2537[a]
R3	61		77.7180[a]
Sig.		1.000	.617

The significant level is set at p=0.05

[a,b] Tukey Post hoc test: groups with the same letter indicate no significant difference. Groups with different letters have a significant difference at the p level.

As the level of residency increases, the mean test score becomes higher. In Figure 6.7, there is a wider range of distribution of scores among first-year residents with most score marks ranging between 60-80%. As the resident's level goes up, their mean test scores shift to the right, with third-year residents having higher test scores (p = 70 – 97%).

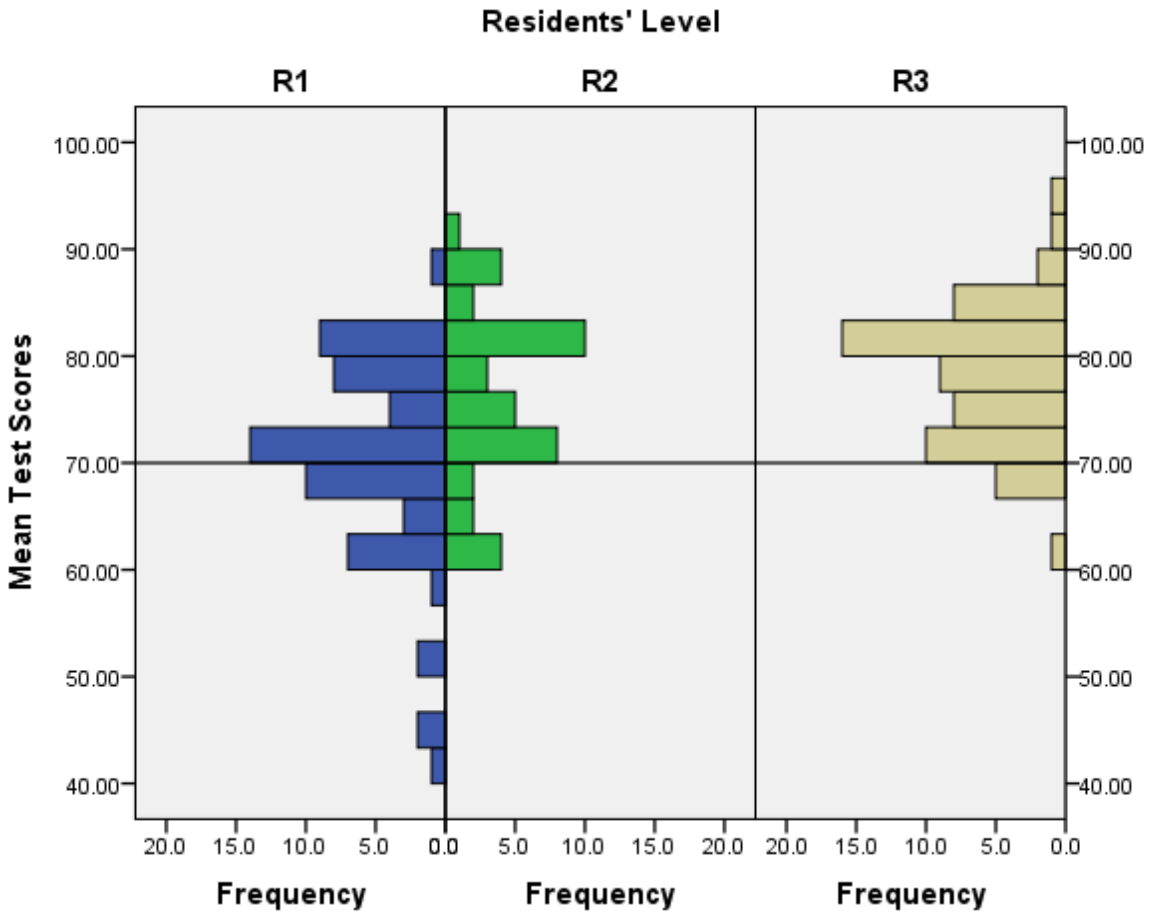


Figure 6.7: Mean test scores according to residency level

Regarding regions, there were no significant differences between mean test scores among residents in the three regions, as shown in Table 6.15 below. Results also showed no association between regions and mean test scores (Eta = 0.076, Eta Sq = (0.01)).

Table 6.15: Mean test scores of pilot (unmarked) items organised by region

Regions	All Groups		Text Group		Multimedia Group	
	N (%)	Mean (SD)	N	Mean (SD)	N	Mean (SD)
Central	112 (68.3)	74.850 (9.01)	55	76.16 (8.02)	57	73.59 (9.79)
Western	26 (15.9)	72.65 (8.16)	15	73.55 (7.85)	11	71.40 (8.79)
Eastern	26 (15.9)	74.33 (7.97)	13	74.09 (9.82)	13	74.56 (5.97)
Total	164	74.42 (8.71)	83	75.36 (8.26)	81	73.45 (9.11)
ANOVA (P-Value)	-	0.511		.470		.689
Eta (Eta Squared)		.076 (0.01)		.137 (.019)		.097 (.009)

6.1.2.2 Combined Data Results of Item Analysis (IA)

Previous results were of exam total scores of the EM residents based on which form they had received (text or multimedia). The focus of this analysis is the comparison of multimedia and text-matched questions. Therefore, the following are results of the parameters of the pilot (unmarked) items only, based on the forms taken. Table 6.16 demonstrates item analysis for the combined data of the unmarked items for the years 2013 and 2015. A paired t-test was done to compare the item parameters: difficulty level (DIFF), discriminating index (DI), point biserial (rPB), and duration to answer the questions between the text (labelled as A) and multimedia items (labelled as B).

Table 6.16: A paired comparison of the means for the psychometric parameters between the multimedia and text items (N=80)

Comparison		Mean	Std. Deviation	Paired Samples t-Test		Paired Samples Correlations	
				t (df)	P value (2-tailed)	Correlation	Sig.
Pair 1[1]	**Diff_A***	.75	.19	1.403 (79)	.16	.811	.000
	Diff_B*	.74	.20				
Pair 2[2]	**DI_A**	.14	.17	-2.160 (79)	.03	.230	.040
	DI_B	.19	.18				
Pair 3[3]	**rPBS_A**	.17	17	-1.676 (79)	.09	.183	.104
	rPBS_B	.21	.19				
Pair 4[4]	**Duration_A**	69.07	21.94	-2.563 (79)	.01	.745	.000
	Duration_B	74.25	26.92				

* A refers to the text items, and B refers to the multimedia items

[1] Diff: refers to difficulty index (easiness of the question) the larger the value the easier the item is and vice versa, it is inversely interpreted.

[2] DI: refers to discriminating index, how much it discriminates between high-ability students and low-ability students, the higher the value the better the discrimination

[3] rPBS: refers to the correlation of the item to the rest of the exam, the higher the value, the more it indicates that the item is testing the same aspect as the rest of the examination.

[4] Duration: time taken in seconds to answer an item.

6.1.2.2.1 Difficulty Index

The mean difficulty index for the 80 text items was slightly higher (i.e., questions were potentially easier) compared to the multimedia items; however, there was no significant difference between both groups (p = .16), which means that no one format was easier or difficult than the other. Both formats had a moderate difficulty level (Diff = 0.75,0.74). The correlation was high between both formats (r=0.81). Put another way, the percentage of students who answered correctly in the text format was not different from the percentage of students who answered correctly in the multimedia format (see Figure 6.8).

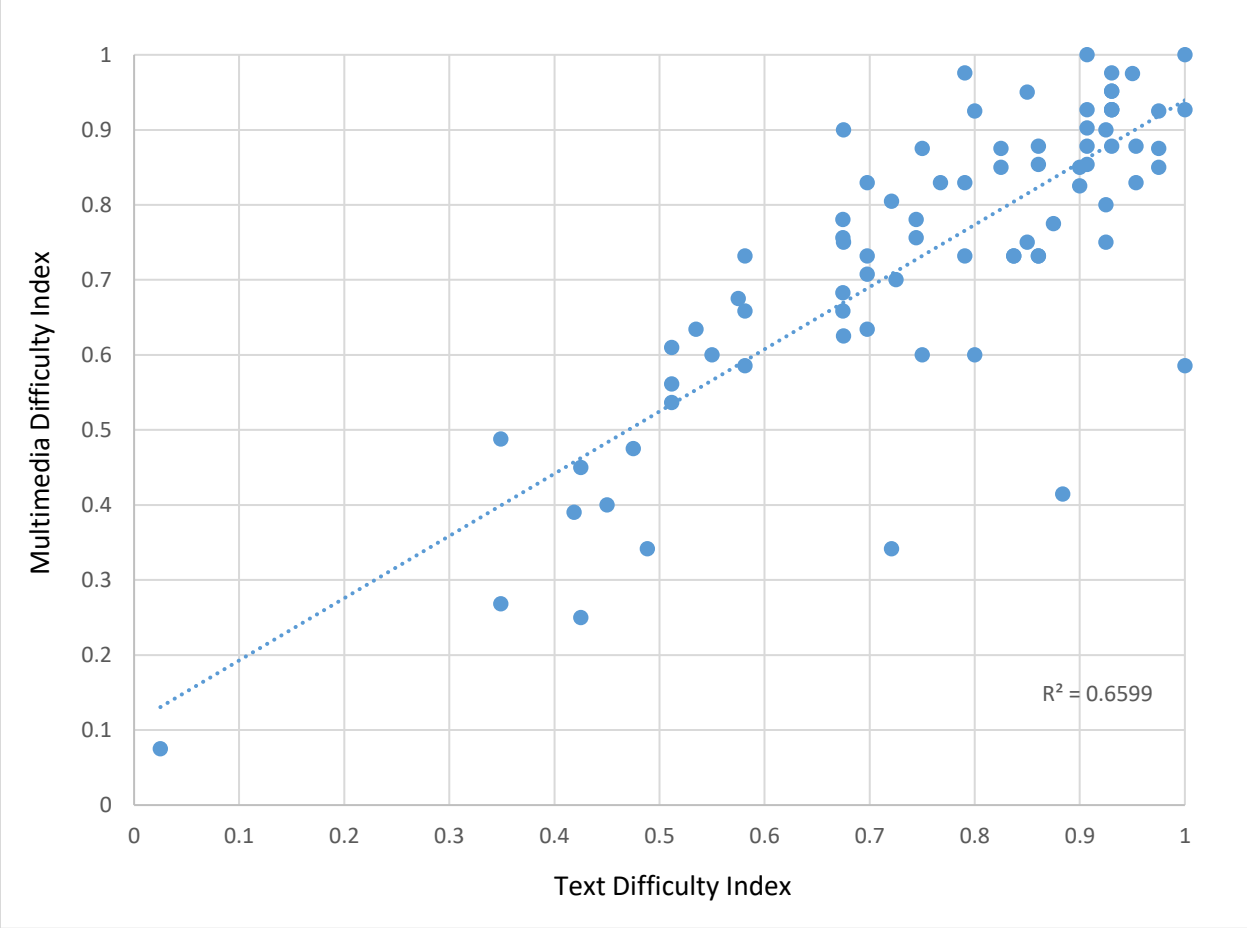


Figure 6.8: Scatterplot of difficulty index between text and multimedia items

6.1.2.2.2 Discriminating Index

The mean discriminating index was significantly lower (and weak) for the text group (DI = 0.14 ±17) than the multimedia group (Diff= 0.19 ±18) (*p*= .03), which shows that the multimedia questions were more discriminating than their text-matched questions. There is a weak correlation (r=0.23) between discriminating indexes for both groups. Items that were discriminating in one format may or may not be discriminating in the other format. This is demonstrated in the scatterplot Figure 6.9. As seen in the scatter plots, the discrimination index between multimedia and text questions have a weak correlation. Items that have high discrimination in multimedia questions may or may not have high discrimination in their text-matched questions and vice versa.

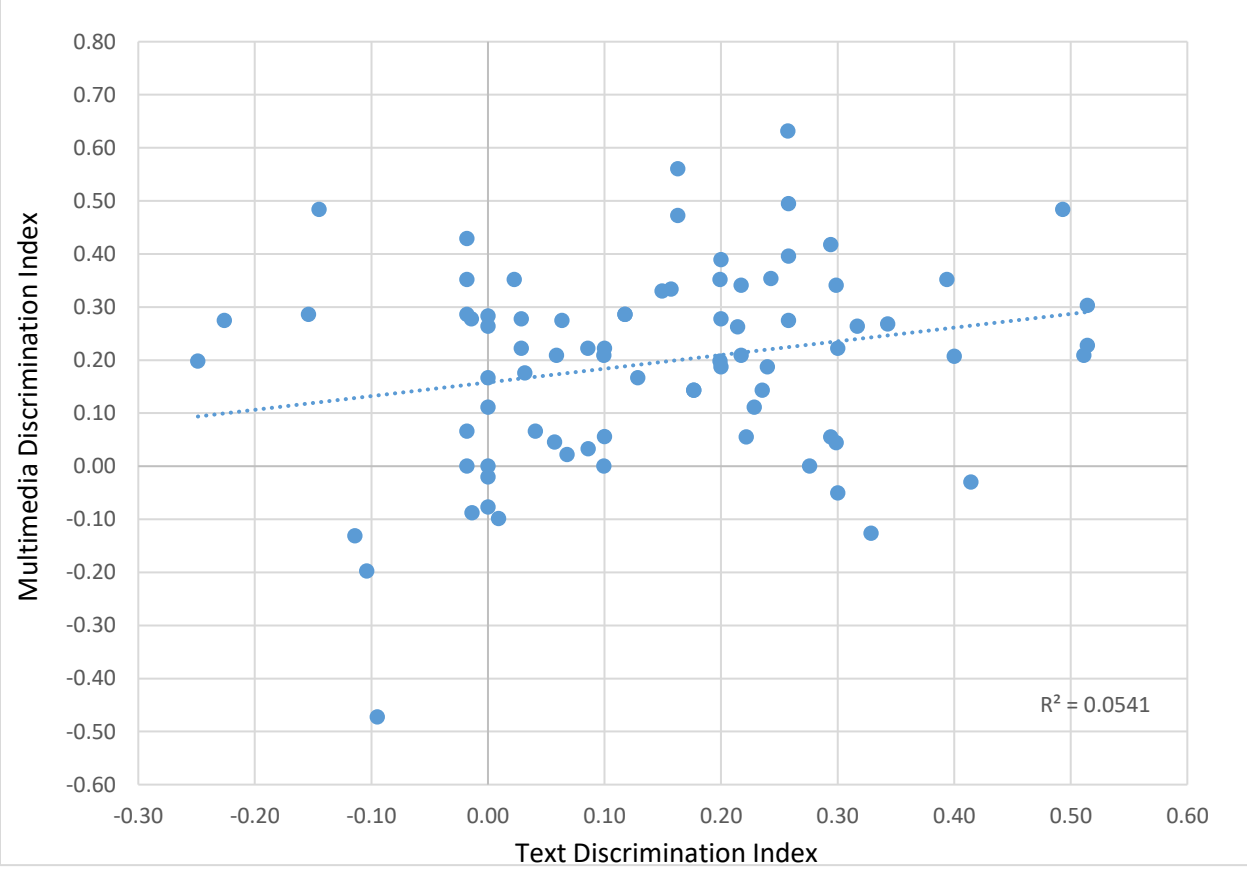


Figure 6.9: Scatterplot of the discriminating index between text and multimedia items

6.1.2.2.3 Point Biserial

The point biserial was higher for the multimedia group than the text group but the p-value did not quite reach significance on a 2-tailed test (p=0.098). It should be noted that although this result was not significant on a two-tailed test, it would have been significant on a corresponding one-tailed test (p=0.049). The two-tailed test could have potentially reached statistical significance with more candidates. The point biserial is related to the discrimination index, and as seen, the higher the discrimination index, the higher the point biserial level is. A high level of point biserial indicates a better correlation of items in a given format to the whole examination. The higher the correlation the more reliable an exam is. Figure 6.10 also reflected a weak correlation (r=0.18) in the scatter plot as did in the discrimination scatter plot.

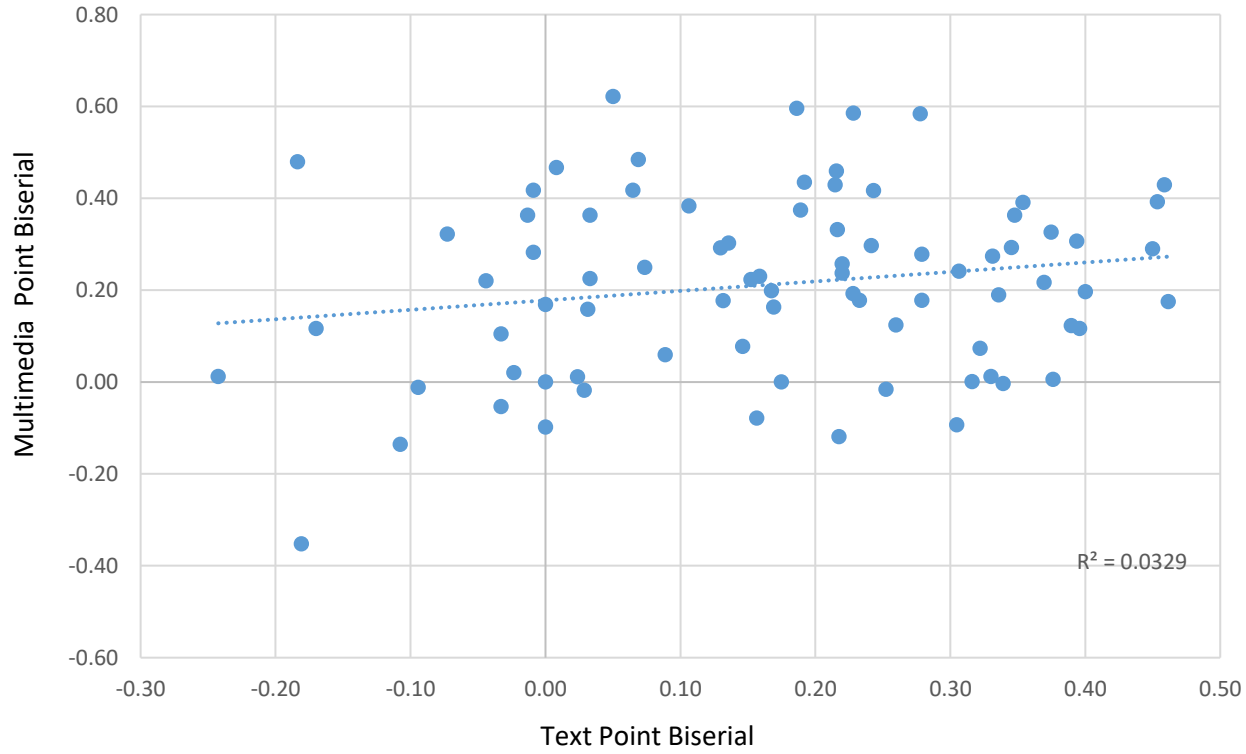


Figure 6.10: Scatterplot of point biserial between text and multimedia items

6.1.2.2.4 Duration

The average time spent on the text-matched questions was (69.07±21.94) seconds, which was less time spent than the multimedia questions (74.25±26.92). The result was statistically significant (p=.01). There was a significantly high correlation between forms (r=.74), which highlights that questions that took longer to answer in the multimedia format also took longer in the text format, as seen in Figure 6.11. Regarding duration, as seen in the scatterplot, there is a moderate to high correlation between text duration and multimedia duration. Overall, questions that had a longer duration in the text format also had a long duration in the multimedia format.

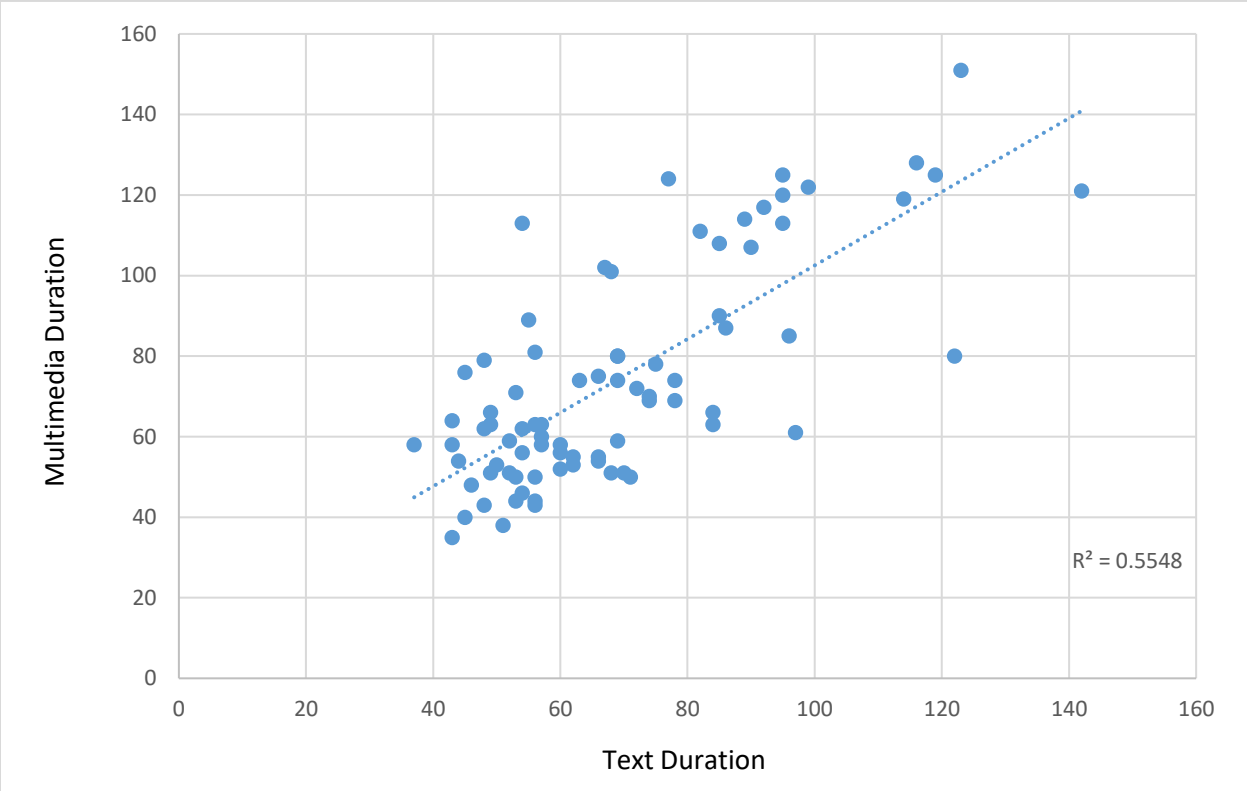


Figure 6.11: Scatterplot of duration between text and multimedia items

6.1.2.3 Correlations

To further understand the relationships between these item parameters, a series of correlations were conducted on both forms using Pearson product-moment correlation coefficient. Regarding the item difficulty (Diff_A) parameter for the text group, items that had a higher difficulty index (i.e., easier) had significantly less time spent on them in order to answer (demonstrated by the -ve correlation r=-.424) in Table 6.17. When the items became easier (higher Difficulty value) the discrimination (DI) became less. (r=-.101) with no correlation. From the correlation, 18% of the durations is explained by the difficulty level (R^2 =0.18). However, the longer the students took to answer a question there was a positive discrimination with a weak correlation (r=.270). From Figure 6.11 a correlation could be seen between the two variables text and multimedia r = 0.74, n = 80, p = 0.000. Items that took longer to answer in the multimedia format also took longer to answer in the text format and vice versa.

As shown in the Table 6.17 below, the multimedia group also had a significant weak to moderate correlation between the difficulty index (Diff_B) and duration (r=-.417), as questions became easier (high difficulty index), the duration to answer the item became shorter. There was no correlation between the discriminating index and the duration of the items as was the case in the text-matched group.

Table 6.17: Correlation between psychometric parameters (Diff, DI and duration) in the text and multimedia groups

Text Group (N=80)		**Duration_A**	**DI_A**
Diff_A	Pearson Correlation	-.424**	-.101
	Sig. (2-taied)	.000	.371
Duration_A	Pearson Correlation	1	.270*
	Sig. (2-taied)	-	.016
Multimedia Group (N=80)		**Duration_B**	**DI_B**
Diff_B	Pearson Correlation	-.417**	.101
	Sig. (2-taied)	.000	.371
Duration_B	Pearson Correlation	1	0.036
	Sig. (2-taied)	-	0.751

**Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

For further analysis, the difficulty index of the items was recoded into three categorical groups (easy, moderate, and difficult), according to the difficulty index values adopted by SCFHS (Table 6.18), and the correlations between parameters were calculated using Pearson's correlation (see Table 6.19).

Table: 6.18: Frequency table for the difficulty level of multimedia and text items (N=80).

Difficulty level	**Diff Index range**	**Text Items Frequency (%)**	**Multimedia Items Frequency (%)**	**Total Frequency (%)**
Difficult	<0.44	6* (7.5)	8* (10)	14 (8.8)
Moderate	0.45-0.79	30 (37.5)	32 (40)	62 (38.8)
Easy	>0.8	44 (55)	40 (50)	84 (52.5)
Total	--	80 (100)	80 (100)	160 (100)

* n is small, and no further calculation can be conducted in this group.

In general, there were a total of fourteen difficult questions throughout the whole MM-TXT format, with 39% medium-difficult questions (n=62) and 53% easy questions (n=84). Except for some questions being slightly more in the multimedia format (n=8), both formats had an approximately equal distribution of moderate and difficult questions.

Because moderate to difficult items (Diff index 0.44-0.79) tend to be more discriminating, the correlation between the moderate-difficulty level and discrimination was calculated for both the multimedia and text-matched items. As seen in Table 6.19, the discrimination was more apparent when we removed the extremely challenging levels of items (easy and difficult questions) in both the text and multimedia groups (when all items were included, r was -0.101 and .101, respectively) as shown in Table 6.17.

Table 6.19: Correlation between psychometric parameters (Diff, and DI) in the moderate difficulty items (0.44-0.79)

Text Group (N=39)		**DI_A**
Diff_A	Pearson Correlation	-.389*
	Sig. (2-tailed)	.037
Multimedia Group (N=40)		**DI_B**
Diff_B	Pearson Correlation	-.113
	Sig. (2-tailed)	.558

*Correlation is significant at the 0.05 level (2-tailed).

In the text group, for questions with moderate-difficulty levels (0.45-0.79), as the difficulty level increased and moved from .45 (difficult) to .079 (moderate) the question became less discriminatory with a significant weak correlation (r= 0.38). There was no correlation between the difficulty level and duration within the moderate group, and it was not looked at here because we would need all the values of easy, as well as difficult items to be included (because duration depends on the items). While in the multimedia group, within the moderate-difficulty questions, there was no correlation between discrimination and difficulty level, as was in the case with the text-matched questions**.**

The following Table 6.20 shows the relationship between difficulty and discrimination index but within the easy items. As the question becomes easier (i.e., the difficulty level increases in value and moves from 0.8 to 1), the discrimination decreases with no significant correlation in the text group and a weak correlation in the multimedia group.

Table 6.20: Correlation between psychometric parameters (Diff, DI and duration) in the easy difficulty items (0.8-1)

Text Group (N=40)		**DI_A**
Diff_A	Pearson Correlation	-.095
	Sig. (2-taied)	.533
Multimedia Group (N=37)		**DI_B**
Diff_B	Pearson Correlation	.133
	Sig. (2-taied)	.385

A Pearson product-moment correlation coefficient was computed to assess the relationship between text and multimedia in their four-item parameters (difficulty, discrimination, point-biserial, and duration). There was a positive correlation between the two variables (text and multimedia difficulty indexes, r = 0.81, n = 80, p = 0.000, and a correlation between text and multimedia discrimination indexes), r = 0.23, n = 80, p = 0.40. A scatterplot summarises the results (Figure 6.8 and 6.9). Overall, there was a strong, positive correlation between difficulty index values of text questions and multimedia questions as demonstrated by having most of the data around the trend line. Items that had a low difficulty index (i.e., hard) in the text questions were also hard (low difficulty index) in the multimedia questions.

6.1.2.4 Crosstabulations

For further analysis, a succession of crosstabulation analysis was conducted to explore the relationships between the different item parameters (difficulty index, discrimination, point bi-serial, and duration) and the different ways of viewing the analysis of these items. In the following sections, an analysis is presented from three main perspectives:

1- Item difficulty level by item analysis (results compiled from residents' exam performance)
2- Item difficulty level by consultants (consultants classified items according to their judgments on how residents would perceive the difficulty level of the item given into easy, medium or difficult)
3- Item cognition level, using the level described in James et al.'s paper (186, 187), the cognitive levels of K1, K2 was used (see Table 5.6, Chapter 5)

6.1.2.4.1 Cross tabs of difficulty level by consultant X difficulty level calculated by IA

All items were classified and reviewed by EM consultants and categorized into easy, medium, and difficult according to their perceptions on how they would expect it to be perceived by the residents. The chi-square was calculated using crosstabs between the perception of the consultant's classification of items into (easy, medium, and difficult) and the difficulty index that was calculated by item analysis of residents' actual performance on the items (into easy, moderate, and difficult) to see if there was any association between them. This was to explore out of those that were classified as easy, medium, and difficult by the consultant, how many actually had a difficulty index of easy, moderate

and difficult (calculated by the residents' performance on the items through the difficulty index) (see Table 6.21 below).

Table 6.21: Crosstabulation-Difficulty level by consultant X difficulty level by item analysis

Difficulty Level (By consultant)[2]		Difficulty_IA[1]			
		Easy IA	**Moderate IA**	**Difficult IA**	**Total**
Easy	**Count**	44	17	4	**65**
	% within Difficulty Level	67.7%	26.2%	6.2%	100.0%
Medium	**Count**	34	38	9	**81**
	% within Difficulty Level	42.0%	46.9%	11.1%	100.0%
Difficult	**Count**	6	7	1	**14**
	% within Difficulty Level	42.9%	50.0%	7.1%	100.0%
Total	**Count**	**84**	**62**	**14**	160
	% within Difficulty Level	52.5%	38.8%	8.8%	100.0%

Chi-square value = 10.384 and p-value = .034
[1] Diff_IA: difficulty level by item analysis (which is calculated by residents' performance on the item)
[2] Difficulty level (by consultant): Perception of the consultants' classification on the difficulty of the items to the residents.

A chi-square test of independence was performed to examine the relationship of item difficulty level between consultants' predictions and resident's performance through IA. The relations between these classifications (consultants' perception and residents' IA) were significant $X^2(4, N = 160) = 10.38$, $p < .03$ indicating a difference between the groups. Consultants were more likely to perceive what would be viewed as easy items by the residents than they would have for moderate and difficult items. In total, consultants had labelled 65 items as being easy, 81 as medium, and 14 as difficult. However, when residents took the exam, 84 items were actually easy, 62 were of moderate difficulty, and 14 were difficult.

Looking at the previous table, the consultant thought that 65 questions would be easy and 81 would be of medium difficulty. However, out of those that they had labelled to be

easy, 44 questions (68%) were actually easy on item analysis (IA), but there were still four questions (6%) that were difficult on the IA. Actually, more items were easy and less were moderate. But interestingly, out of the 14 questions that the consultant thought would be difficult, only one turned out to be difficult. Although the number of classified difficult items by consultants was equal in both groups, the item performance themselves were not. Out of the 14 questions labelled difficult by the consultant, only one was difficult by IA, the rest were almost equally divided between moderate and easy items through the residents' exam performance. From the items that were actually difficult for the residents, four were labelled as easy by the consultant, and nine were labelled as being of medium difficulty. Put in another way, 68% of the consultants' judgement on easy items was matching IA results of residents. But when it came to medium-difficult questions, only 47% of the consultants' judgments was matching the IA results of residents.

6.1.2.4.2 Crosstabulation: difficulty level by consultant X item cognition level by Scientific and Technical Reviewer

In addition to items being classified by the consultants into easy, medium, and difficult, the items have also been classified by the scientific and technical reviewers according to cognition level using the classification that SCFHS has adopted (see Table 5.6, Chapter 5). A chi-square test was performed to examine the relation of consultant's classification of difficulty level and the reviewers' classification of cognitive levels of the items. The relations between these classifications (consultants' perception and reviewers' item cognition labelling) were significant $X^2(2, N = 160) = 6.404$, $p < 0.04$.

Table 6.22 below demonstrates that the consultants' perception of difficulty was significantly different from the cognitive level classification by the reviewers. The

consultant labelled 65 questions as being easy, and in the cognitive classification by reviewers, 43 questions (66%) were of the recall cognitive type (K2-C), which was acceptable. However, out of the 81 questions that the consultant perceived would be of medium difficulty level to the residents, almost half (49%) turned out to be of recall type questions (K2-C) and the other half (51%) were of higher cognitive levels questions (K2-A/B). And out of what was perceived by the consultants to be difficult items to residents (14 items), only five items (36 %) were of recall type questions, and none were of the K2-A type. In fact, two of the K2-A items were labelled as easy by the consultant.

Table 6.22: Crosstabulation: Difficulty Level by consultant X item cognition level by reviewers

Difficulty Level classified by the consultant		Item Cognition Level[1,2] K2-A/B2 (Higher cognitive)	K2-C (Recall)	Total
Easy	**Count**	22	43	**65**
	% within Difficulty Level	33.8%	66.2%	100.0%
Medium	**Count**	41	40	**81**
	% within Difficulty Level	50.6%	49.4%	100.0%
Difficult	**Count**	9	5	**14**
	% within Difficulty Level	64.3%	35.7%	100.0%
Total	**Count**	**72**	**88**	160
	% within Difficulty Level	45.0%	55.0%	100.0%

[1] For simplicity, K2-A/B items are considered items that need analysis and interpretations, and K2-C are items that require recall and understanding.
[2] Observed values were a few in the first column (K2-A) so the data was merged (K2-A/B).
Chi-square value = 6.404 and p value = 0.04

Further, a cross-tabulation was done between item cognition levels and difficulty levels by item analysis (from students' exam performance), see Table 6.23. This was done to examine the relation between reviewers' classification of items' cognitive level and the actual results performed by the residents on the items, which was calculated by IA. There was no difference between the two groups X^2 (2, N = 160) = 1.289, p <.52. The labelling of item cognition level matched and seemed to be aligned with the residents'

performance. Out of the 72 items that were labelled to have a cognition level of K2-A/B, 35 items (49%) of them were found to be easy on IA, 40% were of moderate difficulty, and 11 % were difficult by IA. There was a total of 88 items that were labelled to be as K2-C (recall), with 56% of them being easy, 37% being of moderate difficulty level, and 7% being difficult.

Table 6.23: Crosstabulation: Item cognition level X difficulty level by item analysis

Item Cognition Level[2]		Difficulty IA[1] Easy IA	Moderate IA	Difficult IA	Total
K2-A/B	**Count**	35	29	8	**72**
	% within Cognition	48.6%	40.3%	11.1%	100.0%
K2-C	**Count**	49	33	6	**88**
	% within Cognition	55.7%	37.5%	6.8%	100.0%
Total	**Count**	**84**	**62**	**14**	160
	% within Cognition	52.5%	38.8%	8.8%	100.0%

Chi-square value = 1.289 and p-value = .524
[1] Diff IA: difficulty level by item analysis (which is calculated by residents' performance on the item)
[2] For simplicity, K2-A/B items are considered items that need analysis and interpretations, and K2-C are items that require recall and understanding.

6.1.2.4.3 Independent Sample T-Tests

Further to the crosstabulation results, an independent sample t-test was conducted to compare item parameters with item cognitions of K2-A/B and K2-C levels.

Table 6.24: Independent samples tests for cognition levels

Item Cognition Level[1]	Difficulty	Discrimination	rPBS	Duration
K2-A/B (Mean ± SD)[2]	.72 (.20)	.18 (.20)	.20 (.20)	81.38 (27.25)
K2-C (Mean ± SD)	.77 (.18)	.15 (.16)	.18 (.16)	63.72 (18.97)
t[3]	-1.631	1.098	.749	4.654
p-value	.105	.274	.455	.000

[1] N for K2-A = 2, K2-B = 70, K2-C = 88.
[2] Omitting K2-A items or combining them yields the same results.
[3] Degree of freedom for the duration was 122 and the rest *df* =158.

As demonstrated in Table 6.24, there was no difference with regards to the difficulty, discrimination or the point biserial (rPBS) in regards to the type of items based on cognition level. The only significant difference was in the duration parameter. Residents took more time to answer K2-A/B questions (M = 81.38, SD = 27.25) than they did on the K2-C items (M = 63.72, SD = 18.97), t (122) = 4.65, p = .000. In order to compare the psychometric parameters between the consultant's difficulty levels, a one-way ANOVA was conducted and is shown in Table 6.25.

Table 6.25: One-way ANOVA of psychometric parameters (Diff, Dis, rPBS, and Duration) by difficulty level perceived by consultant

Difficulty level by consultant (Mean ± SD)	**Difficulty**	**Discrimination**	**rPBS**	**Duration**
Easy (N =65)	.81 (.16)	.13 (.15)	.18 (.19)	65.72 (22.80)
Medium (N =81)	.70 (.21)	.20 (.19)	.19 (.18)	74.42 (25.58)
Difficult (N =14)	.70 (.17)	.18 (.17)	.22 (.19)	83.29 (21.38)
F[1]	6.182	3.062	.289	4.116
p-value	.003	.050	.749	.018

[1] the degree of freedom = 2

The 65 questions that the consultants had said would be perceived as easy items by the residents had a mean value of .81 ± (.16). Item analysis had an equal value for both moderate and difficult items (X = 0.70). Easy items had a lower discrimination index as compared to the medium and difficult items, with medium items having the highest discrimination index. Regarding the duration, the more difficult the item was, the more time was needed to answer the questions. An analysis of variance showed that the effect of difficulty was significant, F (2,160) = 6.182, p = 0.003. Post-hoc analysis using the Tukey post-hoc criterion for significance (alpha = 0.05) indicated that the average mean

for the difficulty index was significantly higher in the easy questions (M = 0.81, SD = 0.16) than in the medium questions (M = 0.70, SD = 0.21), p = 0.003.

Also, there was a significant difference in the duration variable F (2,160) = 4.116, p = .018. Using the post-hoc test (Tukey's test). The difficult items as marked by the consultant had a longer duration to answer (M = 83.29, SD = 21.38) as compared to the easy items (M = 65.72, SD = 22.80), p = 0.04.

Regarding discrimination, although it demonstrated a borderline significant difference F (2,160) = 3.062, p = .050, on doing a Tukey's post hoc test, it was found that there was a significant difference in discrimination. Medium difficulty items had a higher discrimination index (M = 0.20, SD = 0.19) than easier items (M = 0.13, SD = 0.15) p= 0.04.

Similar results were not achieved for the difficult questions, because the power was affected with the low count (14 items). Therefore, these items were recoded with the medium items and a t-test was recalculated splitting the data by their two forms (multimedia, and text) to better understand the results, as the previous results were to understand the items as a whole set (160 items) rather than by their forms.

6.1.2.4.3.1 Results of (t-tests and crosstabs) by forms

For both the multimedia and text groups, the following Table 6.26 displays the t-test and means for the four IA parameters by difficulty levels that was assigned by the consultant. Here, by taking a quick look at the table above in the text items, items that were labelled by the consultant as easy were actually easier on IA with an average of 0.80, and those that were labelled to be of medium/difficult were harder according to IA (mean 0.70), with a p-value = 0.02. There was a difference between the discrimination of the easy items DI=0.11 and the rest of the items DI= 0.17 in the text group, with a p-value of borderline significance. There were no differences regarding the point biserial (rPBS) and duration in the text questions.

Table 6.26: T-test for item difficulty level by consultant (regrouped) X IA parameters

	Text Items (N for Easy = 42, Med/Diff = 38) [1]				**Multimedia Items (N for Easy = 23, Med/Dif = 57)**			
Item Difficulty Level[2]	**Difficulty**	**Discrimination**	**rPBS**	**Duration**	**Difficulty**	**Discrimination**	**rPBS**	**Duration**
Easy	.80 (.17)	.11 (.16)	.14 (.17)	65.57 (22.19)	.82 (.14)	.15 (.14)	.25 (.20)	66 (24.38)
Med/Diff	.70 (.20)	.17 (.17)	.20 (.17)	72.95 (21.28)	.70 (.21)	.21 (.20)	.20 (.19)	77.58 (27.37)
t[2]	2.354	-1.711	-1.358	-1.514	2.987	-1.229	1.165	-1.764
p-value (2-tailed)	.02	.09 *	.18	.13	.004	.223	.248	.082**

[1] means and (standard deviations) are presented in the cells.
[2] Difficulty level by consultant after recoding Difficult items with medium ones.
[3] Degree of freedom for Text and MM DF = 78, for MM Diff DF = 59.19
* Border line significant on 2-sided t-test, border-line significant on one-sided t-test (p=0.046)
** Border-line significant on one-sided t-test (p=0.041) for one-sided hypothesis

Reviewing the multimedia (MM) group of items only, there was a difference in item difficulty and duration but not within point-biserial or discrimination values. In the MM items, the difficulty level from the IA is quite the same as that of the text group in both the easy and medium/difficult questions that were labelled by the consultant. The question was difficult no matter what the format was. Discrimination was higher in the MM items, particularly in the medium/difficult group however, it was not significant.

Tale 6.27: T-test for Item Cognition level by reviewers X IA parameters[1].

Item Cognition Level	Text Items (N for K2-B = 13, K2-C = 67) [1]				Multimedia Items (N for K2-B = 59, K2-C = 21)			
	Difficulty	**Discrimination**	**rPBS**	**Duration**	**Difficulty**	**Discrimination**	**rPBS**	**Duration**
K2-A/B	.74 (.17)	.21 (.22)	.23 (.20)	81.15 (28.58)	.72 (.22)	.18 (.20)	.20 (.21)	81.42 (27.20)
K-2C	.76 (.20)	.12 (.15)	.16 (.16)	66.73 (19.84)	.80 (.10)	.24 (.15)	.25 (.15)	54.10 (11.76)
t[1]	-.303	1.758	1.316	2.222	-2.426	-1.299	-1.152	6.249
p-value (2-tailed)	.763	.083**	.192	.029	.018	0.198	0.253	0.000

[1] Omitting K2-A items or combining them yields the same results.
[2] Degree of freedom for Text and MM DF = 78, for MM DF = 76, DF for MM Diff= 71.377, DF for MM Duration = 74.949
** Border line significant on 2-sided t-test, borderline significant on one-sided t-test (p=0.041)

Data were explored by item cognition level; an independent sample t-test was carried out on each form to compare the means of higher cognitive items (K2-A/B) and recall items (K2-C) (see Table 6.27). Within the text format of items, there was a borderline significance in the discrimination index between the recall (K2-C) items and the higher cognitive level items (K2-A/B), with K2-A/B having a higher discrimination level. In addition, K2-A/B items took more time to answer than the K2-C items (p = .029). In the multimedia group of items, there is a significant difference in the difficulty level index between K2-A/B items and K2-C items (p = 0.18). K2-A/B items took longer to answer than K-2C items with a p-value of 0.000. Equal time was spent on both the multimedia and text items in the K2-A/B items (duration = 81 seconds). Because items in the difficult category were minimum, difficult items were recoded to be combined with moderate items and an independent sample t-test was again used to compare the psychometric parameters (Diff, Dis, rPBS and duration) on the basis of difficulty level by item analysis when the groups were merged into two groups (easy and moderate/difficult).

Table 6.28: Independent sample T-test for difficulty levels by IA X IA parameters[1]

	Text Items (N for Easy items = 44, Moderate/Difficult = 36)				Multimedia Items (N for Easy items = 40, Moderate/Difficult = 40)			
Difficulty by IA into (two groups)	**Difficulty**	**Discrimination**	**rPBS**	**Duration**	**Difficulty**	**Discrimination**	**rPBS**	**Duration**
Easy IA	.90 (.06)	.11 (.14)	.17 (.18)	61.82 (18.79)	.89 (.06)	.19 (.15)	.28 (.20)	65.55 (22.73)
Moderate/Diff IA	.58 (.15)	.18 (.19)	.17 (.16)	77.94 (22.49)	.59 (.17)	.19 (.22)	.14 (.17)	82.95 (28.21)
t[1] (df)	11.451	-1.775	.125	-3.494	10.704	-.036	3.316	-3.038
p-value (2-tailed)	.000	.081*	.901	.001	.000	.971	.001	.003

* Border line significant on a 2-tailed test and (significant p=0.04) on a 1-tailed test

[1] *df* for text (diff = (44.497), DI = (61.838), rPBS and duration = (78)); for MM (diff = (49.55), for DI, rPBS and duration = (78)).

Table 6.28 demonstrates the t-test for item difficulty levels by IA and in both the text and multimedia format. The difficulty level and duration of the items were significantly different between the easy items and the moderate/difficult items. Items in both formats had approximately the same difficulty levels in both subgroups. Easier items took less time to answer than moderate/difficult items in both formats. Regarding the discrimination index in the text format, there was a borderline significance between easy items and moderate/difficult items, with moderate/difficult items being more discriminating (0.18$\pm$0.19) than the easy questions in the text group (0.11$\pm$0.14). Within the MM group, there was no discrimination between the easy and moderate/difficult items, as was in the text group; however, it is noticed that easy items in the multimedia group had higher discrimination than in the text group. The point biserial in the MM format was significantly higher in the easy items (M = 0.28, SD = .20) than in the moderate/difficult items (M = .14, SD = .17). Residents took a longer time to answer the multimedia questions in both the easy and moderate/difficult items than they did in the text format.

6.1.2.4.4 Crosstabulation of exam formats by item cognition level

Finally, because the nature of the question changes when an image or video is added to it, the level of cognition and difficulty level might be affected. Therefore, a series of crosstabulations have been conducted to compare the results by exam format (text or multimedia) with item cognition level, item difficulty level perceived by the consultants, and item difficulty levels through residents' performance (item analysis). First of all, a cross-tabulation between forms and item cognition levels is presented in Table 6.29.

Table 6.29: Crosstabulation of forms X cognition

		Item Cognition Level		
Form		**K2-B/A**	**K2-C**	**Total**
TXT	**Count**	13	67	**80**
	% within form	16.3%	83.8%	100.0%
MM	**Count**	59	21	**80**
	% within form	73.8%	26.3%	100.0%
Total	**Count**	**72**	**88**	160
	% within form	45.0%	55.0%	100.0%

*Chi-square test value =53.43, df,1 p-value = .000

When running a crosstab between item format and item cognition level, there was a difference between the multimedia (MM) and the text (TXT) format in regards to the items' cognitive level, X^2 (1, N = 160) = 53.43, p =.000. It was found that the majority of items in the text format were of the recall/understanding type (K2-C = 84%) while the majority of the multimedia items were of the higher cognitive levels (K2-A/B = 74%).

Within the text group only, 16% of the items had a cognitive level of K2-A/B labelled by the reviewers, and most items (84%) had been labelled as recall (K2-C) while the opposite was noticed in the MM group, with 74% of the items thought to be of higher cognitive level K2-A/B by the reviewers and only 26% were thought to carry a recall level of cognition.

Next, a crosstabulation between forms and consultants' views on how the items would be viewed by the residents were carried out and results are shown in Table 6.30 below.

Table 6.30: Crosstabulation of form X difficulty level by consultant

Form		Difficulty Level by consultant Easy	Medium / Difficult	Total
TXT	**Count**	42	38	**80**
	% within form	52.5%	47.5%	100.0%
MM	**Count**	23	57	**80**
	% within form	28.8%	71.3%	100.0%
Total	**Count**	**65**	**95**	160
	% within form	40.6%	59.4%	100.0%

*Chi-square test value =9.35, df =1, p-value = .002

The table demonstrates similar results as the cognition level (see Table 6.29 above). Results show that there is a significant difference between forms according to difficulty level classified by the consultant X^2 (1, N = 160) = 9.35, p =.002). In the text format, according to the consultant's labelling, 53% of the items were easy and 48% were of medium/difficult level while in the MM form, 71% of the items were classified as medium/difficult and 29% as easy. The text format had almost equal distribution of easy and medium/difficult items while the multimedia format comprised mainly of medium/difficult items as labelled by the consultant.

Lastly, a crosstabulation between item cognitive level and item analysis from performance for each of the format is presented. Both crosstabs in tables 6.31 and 6.32 show that there was no association between item cognitive levels and the difficulty level by IA. There was no difference between the two groups in the text format, as well as in the multimedia format.

Table 6.31: Crosstabulation: item cognitive level X difficulty by IA (text format)

Item Cognition Level		Difficulty by IA Easy	Moderate/Difficult	Total
K2-B/A	**Count**	6	7	**13**
	% within form	46.2%	53.8%	100.0%
K2-C	**Count**	38	29	**67**
	% within form	56.7%	43.3%	100.0%
Total	**Count**	**44**	**36**	80
	% within form	55.0%	45.0%	100.0%

*Chi -square test value =.491, df =1, p-value = .484

Table 6.32: Crosstabulation: Item cognitive level X difficulty by IA (multimedia format)

Item Cognition Level		Difficulty by IA Easy	Moderate/Difficult	Total
K2-B/A	**Count**	29	30	**59**
	% within form	49.2%	50.8%	100.0%
K2-C	**Count**	11	10	**21**
	% within form	52.4%	47.6%	100.0%
Total	**Count**	**40**	**40**	80
	% within form	50.0%	50.0%	100.0%

*Chi -square test value =0.065, df =1, p-value = .799

6.1.2.5 Reliability (Cronbach Alpha)

Reliability was calculated using SPSS V24. Cronbach's Alpha for the promotion exam and multimedia text-matched items were calculated for each year. Tables 6.33 and 6.34 demonstrate the results.

Table 6.33: Reliability for 2013 items using Cronbach's Alpha

	100 Standard	30 Items	130 Items
Text Group	.765	.096	.761
Multimedia Group	.810	.405	.835

Reliability results for the promotion exam alone (100 questions) for the text group in 2013 was (α = 0.765). Adding the text items that were taken by the residents slightly decreased the exam reliability to (α = 0.761). However, multimedia items had the opposite effect. When adding it to the promotion items, it increased exam reliability. When calculating results for the multimedia and text items reliability alone they were low as the number of items was much less (30 items in 2013 and 50 items in 2015). However, in 2013, the multimedia items had a higher reliability than the text items as seen in Table 6.33.

Items for the year 2015 had a higher reliability value than 2013 due to the higher number of items. Table 6.34 also shows that taking pilot items alone, text items had slightly higher reliability (α = .636) than multimedia items (α = .615). However, generally, the multimedia groups had higher reliability.

Table 6.34: Reliability for 2015 items using Cronbach's Alpha

	100 Standard	50 Items	150 Items
Text Group	.797	.636	.849
Multimedia Group	.844	.615	.879

6.1.2.6 G-Coefficient and D-Study

Each student took the 100 promotion questions and one format (either MM or TXT). So, in this research, the promotion questions were crossed as each student took the promotion items. In addition, students were nested in forms (students: forms) as each student could only take one form in the exam (either MM or TXT). MM and TXT items were also nested in forms (items: form) since each form had its unique items. The possible sources of error to be studied here were items, forms, and person and are considered as random facets. Levels, regions, and gender are fixed facets (193). The G-coefficient was calculated using Excel 2016 and SPSS V24. The variance was analysed using the MINQUE technique to measure the sources of variances contributing to results that could not be measured by using the classical test theory reliability (Cronbach).

Table 6.35 displays the variance component estimates from the generalizability study. There are two contributors to errors in this research, students and items, as well as errors that are not accounted for. For the multimedia examination that was run in 2013 (pilot), the student variance component accounted for 1.7 % of the total variance. Indicating that differences among students' characteristics were minimal and didn't vary substantially overall. The multimedia items accounted for 22.6 % of the total variance. This indicated that the multimedia items varied in characteristics and parameters. Almost 76% of the variance was unaccounted for (due to error) and could be explained by the students' interaction with the items. The G-coefficient was .40 and is considered low; however, it is based only on 30 items. The more good quality items are involved in a test the higher the reliability is expected to be.

In the text examination for 2013 (pilot) 24% of the variance was due to the items indicating variation among the text characteristics, and the rest of the variances was attributed to error (or students' interaction with the item). Examination in 2015 for both formats (text and multimedia) items accounted for almost 15% of the total variance and almost 3% were due to student variances. The rest of the variances were due to error or interaction of the students with the items. The G-coefficient (another measurement of reliability) was higher for the multimedia items in 2013 than the text and similar to text in 2015 items. The G-coefficient was increased for both formats in 2015 and indicated that a G-coefficient of .62 for the multimedia items.

Table 6.35: G-Coefficient and sources of error variance for multimedia and text

Source	**(Con %) * MM 2013**	**(Con %) TXT 2013**	**(Con %) MM 2015**	**(Con %) TXT 2015**
Student	1.7	0.3	2.7	2.9
Item	22.6	24.1	15.8	14.5
Error	75.7	75.6	81.5	82.6
Items (n)	30	30	50	50
G-Coefficient	.40	.10	.62	.64
G SEM**	7	7	6	6

*Con refers to contribution expressed in percentage
** Standard Error of Measurement

The SEM for each of these exams was calculated and was found to be seven for both exams forms in 2013 and six for both exams forms in 2015. The SEM was used to calculate a 95% confidence interval (by multiplying it by 1.96) and resulted in a confidence interval of ±13.72 for the 2013 exams and ±11.76 for the 2015 exams.

A decision study (D-study) based on the results of the G-coefficients was generated and the results of items needed are displayed in Table 6.36. The table displays both the G-coefficient, as well as phi-coefficient of the D-study at 0.80 (the desired level of exam reliability for high-stakes exams). Phi-coefficients tend to be lower than G-coefficients and

results are calculated based on the assumption of a new exam with new items and new participants. In Table 6.36, the D-study indicated that a total of 122 multimedia items would be required to achieve a G-coefficient of 0.80 based on the 50 multimedia items that were used in 2015, and would require 144 new multimedia items to reach a phi-coefficient of 0.80. It can be seen from the table that the estimated number of items needed based on the 2015 exam is less than those needed based on the 2013 exam. In addition, based on the text items of 2015, a D-study indicates that 111 of the same text items would be required to reach a G-coefficient of .80 and 131 new text items would be required to achieve a phi-coefficient of 0.80.

Table 6.36: D-Study for multimedia and text based on a desired coefficient of 0.80

Coefficients (set at 0.80)	**G[1]**	**Phi[2]**
Items needed based on MM 2013	171	222
Items needed based on TXT 2013	1168	1450
Items needed based on MM 2015	122	144
Items needed based on TXT 2015	111	131

[1] G-Coefficient for the generalizability study
[2] Phi- Coefficient for the D-study

When comparing reliability calculated from the G-theory results with Cronbach's alpha reliability for the multimedia and text items of both years results were similar. Table 6.37 outlines the results.

Table 6.37: Reliability results of Cronbach alpha and G-coefficient

Items (year)	**30 Items (2013)**		**50 Items (2015)**	
Reliability	**Alpha**	**G-Coefficient**	**Alpha**	**G-Coefficient**
Text Group	.09	.10	.63	.64
Multimedia Group	.40	.40	.61	.62

6.2 Questionnaire (survey) results

This report outlines the results of the online survey that was conducted to discover and explore the opinions of residents towards computer-based testing and the use of multimedia (image and video-based) questions in their examinations. The survey consisted of 50 statements in six main themes, rated on a four-point Likert scale (Appendix 11). Results were calculated using IBM SPSS V.24 and Microsoft Excel 2016 to analyse the results displayed through frequency tables and bar charts. The items were coded and the four-point Likert scale was given ordinal values from 1-4.

There were 84 residents who had access to the questionnaire, with 69 of them (82%) starting the survey. 23% (n=19) discontinued the survey from the first theme. The final response rate, therefore, was 59% (n= 50). Out of these 69 residents, almost half were from the text group and the other half were from the multimedia group. Most of the residents who completed the survey were from the Central Region, followed by the Western Region, then the Eastern Region. Regarding gender, males were more than females by three-fold as seen in Table 6.38. Residents who answered the survey were from all levels with equal distribution from the junior levels (R1 and R2) and more being from the senior levels (R3).

Table 6.38: Frequency of residents' demographic completing the survey

Form	Gender	Region	Level
TXT = 32	Male = 63	Central = 48	R1 = 17
MM =37	Female = 21	Western = 13	R2 = 20
-	-	Eastern = 8	R3 = 32

The questionnaire contained a few questions that were not applicable to the text group to answer and were only for those who received the MM items. The survey contained some item-non-response from some participants. However, their responses on a complete theme were included; otherwise, they were removed from the rest of the survey for failure to complete the survey.

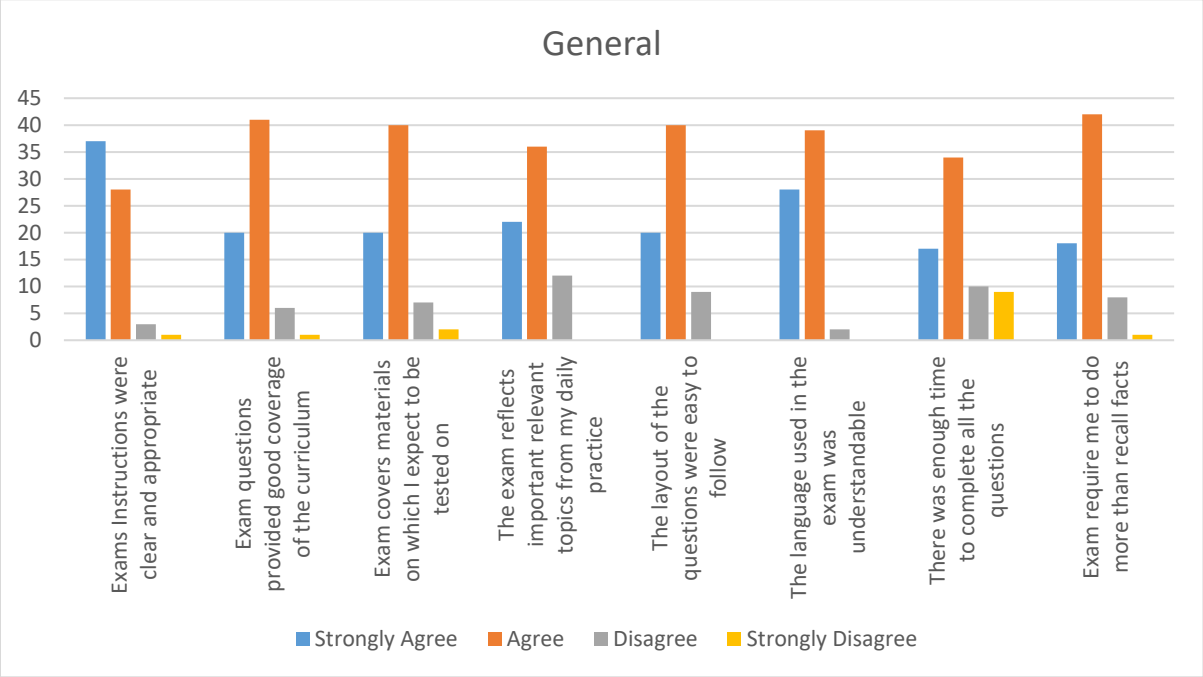


Figure 6.12: Survey results for the general questions (Theme 1)

As seen from the graph above (Figure6.12), results from the six general theme questions, all of the items in this theme favoured the agreement. The highest disagreement was on not having enough time to complete all the questions, with 27% of the resident's selecting either disagree or strongly disagree (this is further explained and elaborated on in the qualitative analysis section). Fifty-four per cent of the residents felt strongly about having their exam instructions clear and appropriate.

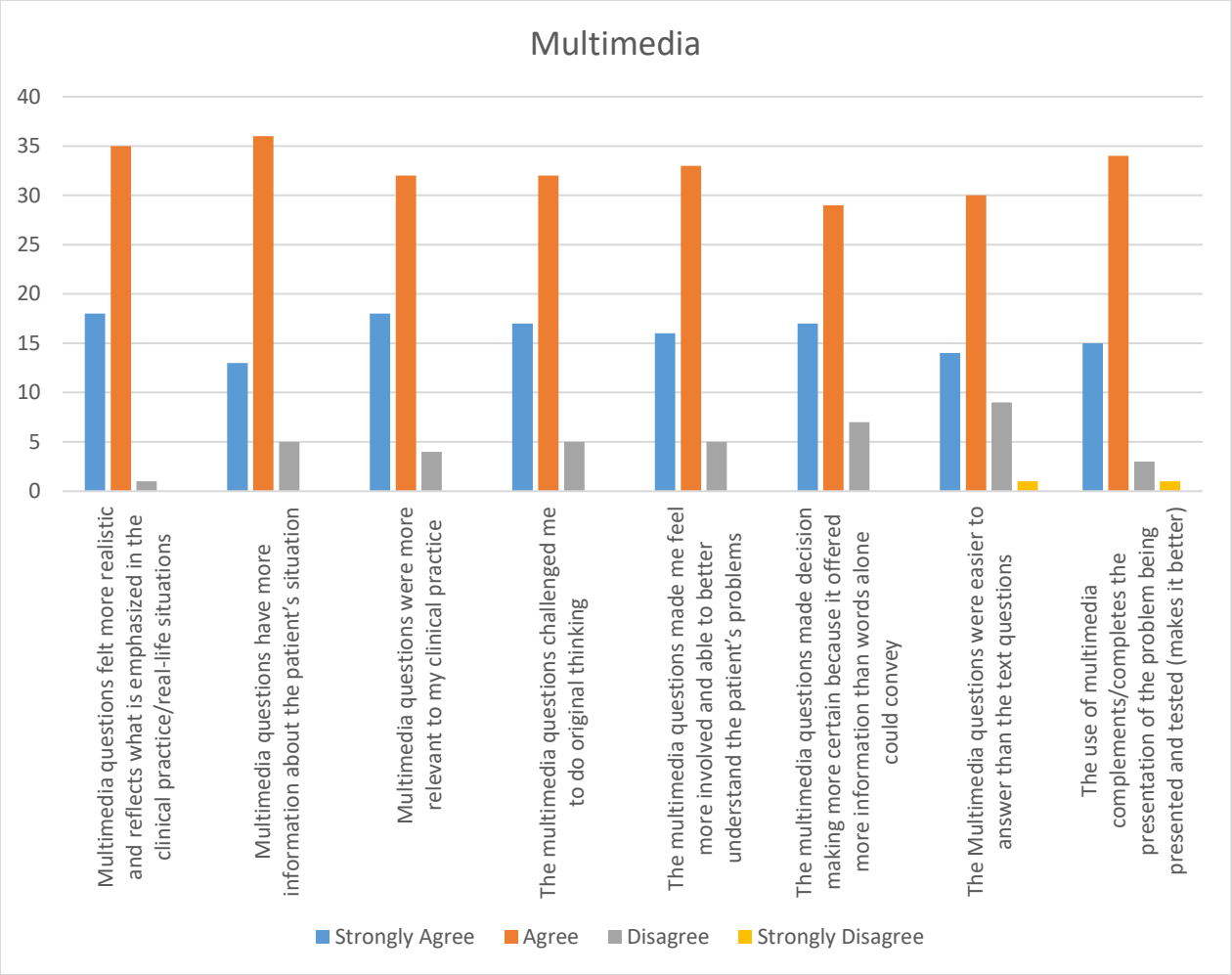


Figure 6.13: Results for multimedia questions (Theme 2)

Regarding the multimedia questions theme (Figure 6.13), the same was observed regarding the eight statements with 32.6% mean average of agreement and 2.6 % of disagreement.

Looking at the results in Figure 6.14, there was a high agreement level among residents regarding the value and use of images in examinations. Some residents felt that the use of image enlargement was not useful and that images didn't necessarily make them do more than recall facts.

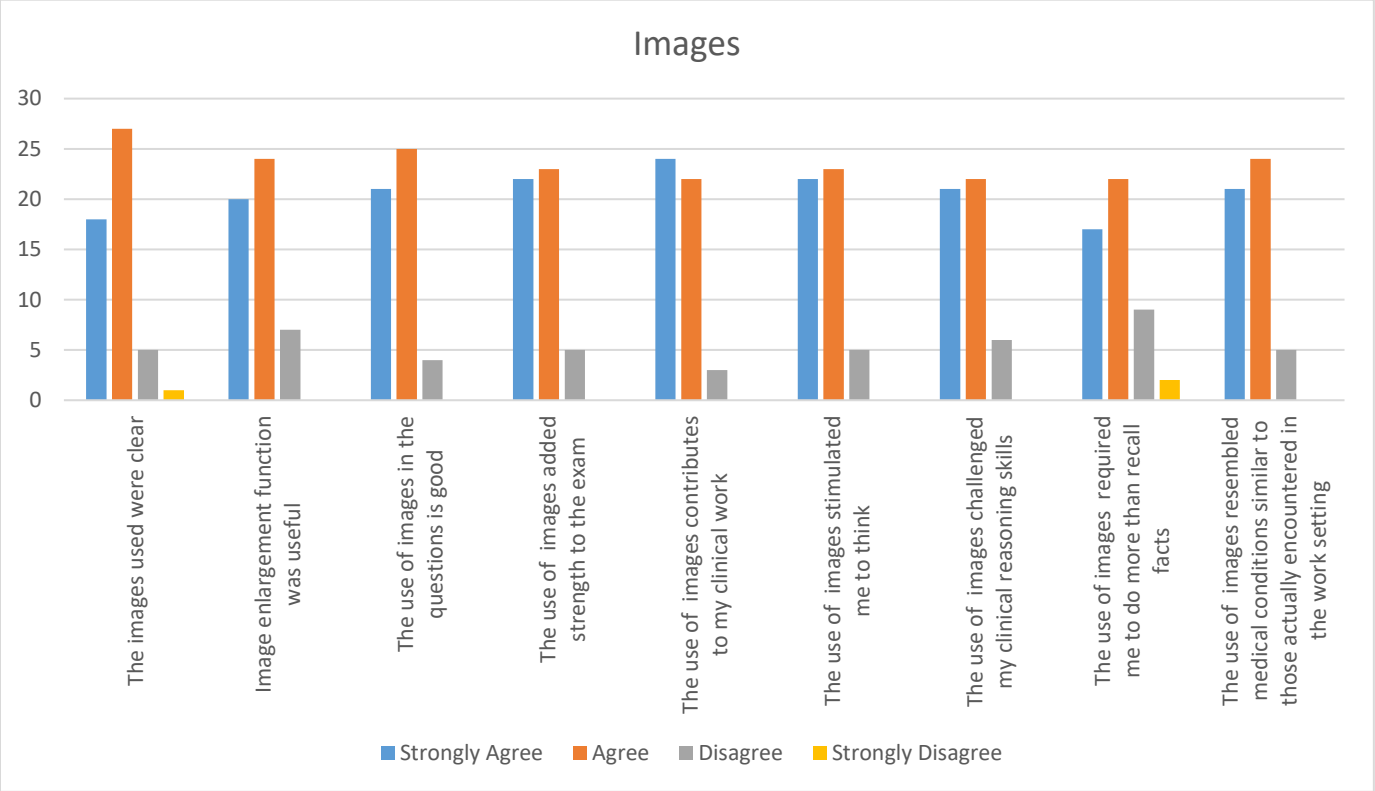


Figure 6.14: Results for images questions (Theme 3)

The same was observed in the questions related to the video theme (figure 6.15) of using videos in examinations and what it entailed. An overall view demonstrated that most of the residents were in high agreement and were in favour of using videos in examinations.

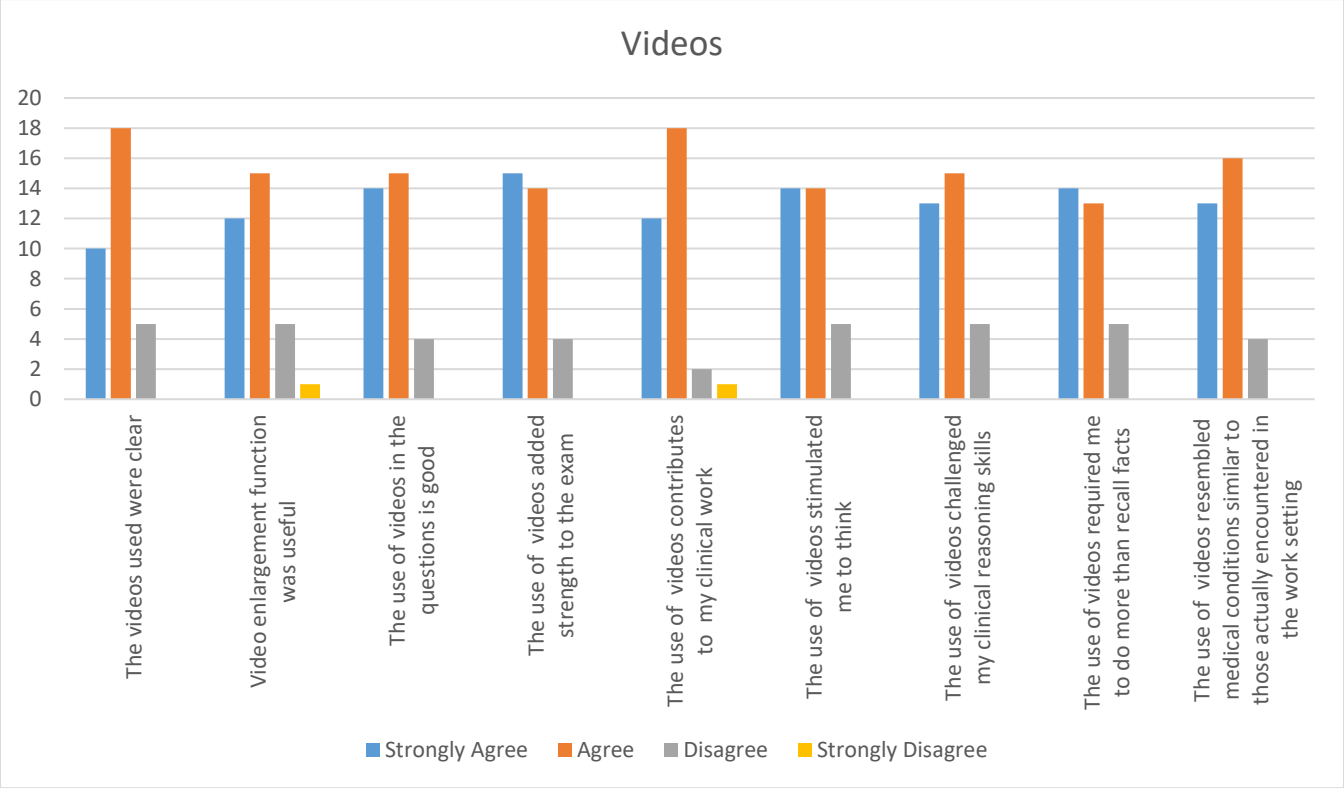


Figure 6.15: Results for video questions (Theme 4)

Regarding computer-based questions, most of the residents who completed the survey agreed that the use of computer-based testing and most of their functions were helpful during their examination. There seemed to be some conflicting opinions regarding one of the CBT functionalities (i.e., scrolling) among residents with 35% disagreeing it was distracting to their examinations and 65% agreeing that they were distracted during their examinations by the scrolling function (see Figure 6.16).

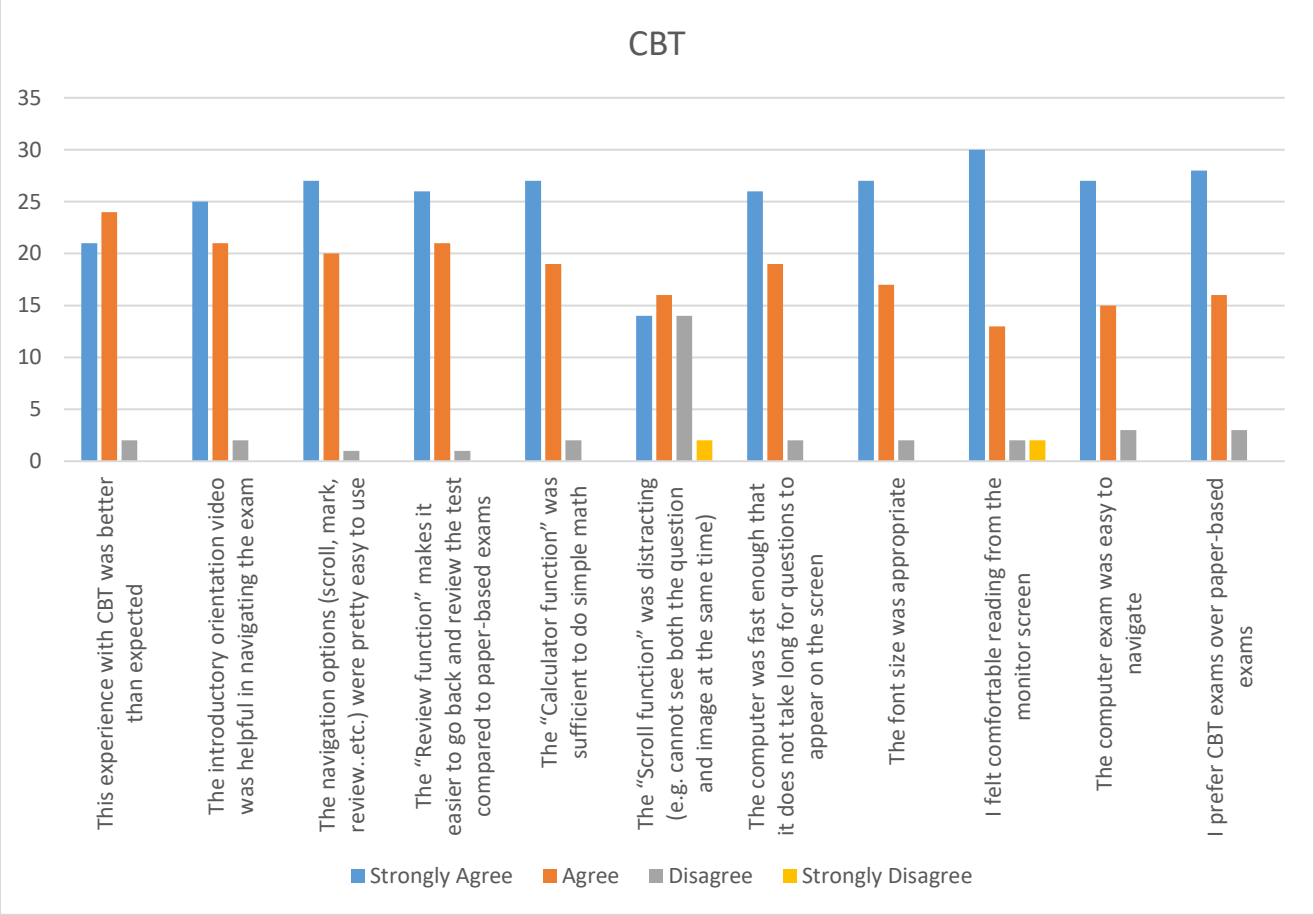


Figure 6.16: Results for computer-based testing questions (Theme 5)

Overall, residents were satisfied with the staff professionalism, environment technical, and overall experience with their examinations, as seen in the bar charts below in Figure 6.17. Some residents felt uncomfortable with regards to room temperature. Figure 6.18

highlighted that most residents didn't face any technical problems during their examinations.

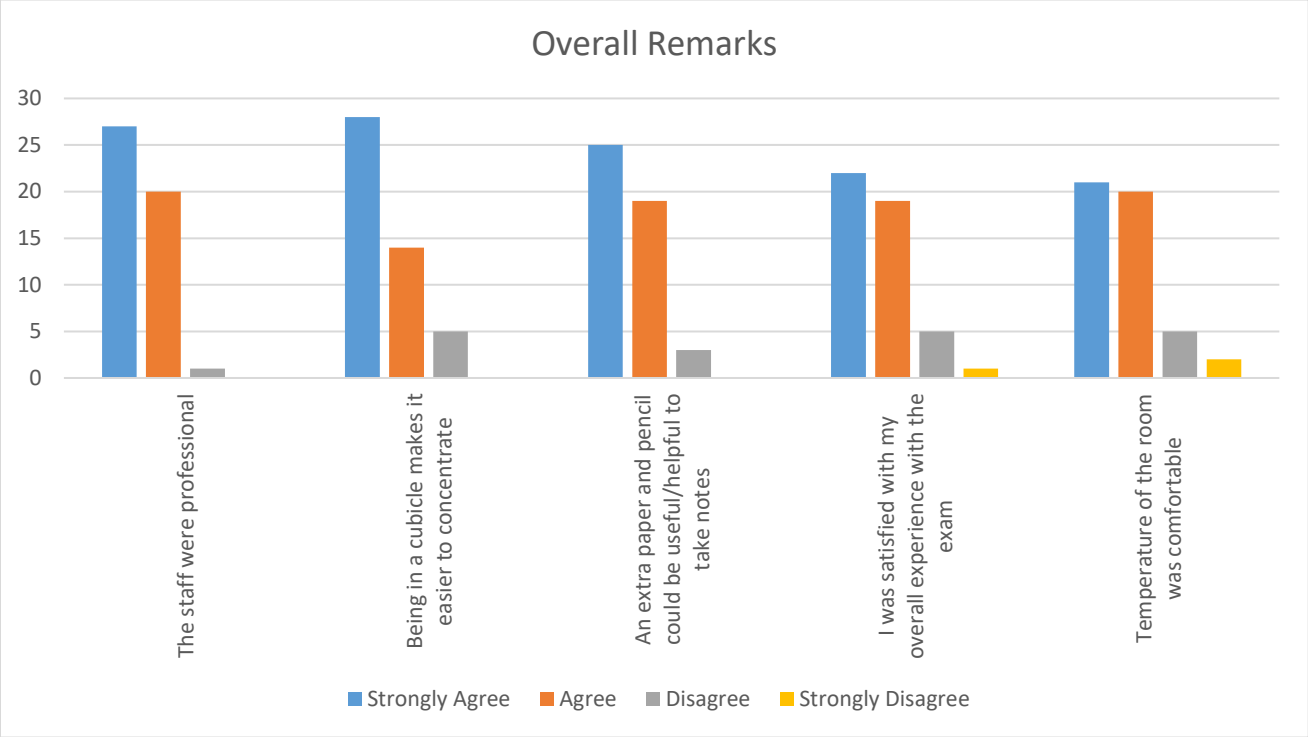


Figure 6.17: Overall results questions (Theme 6)

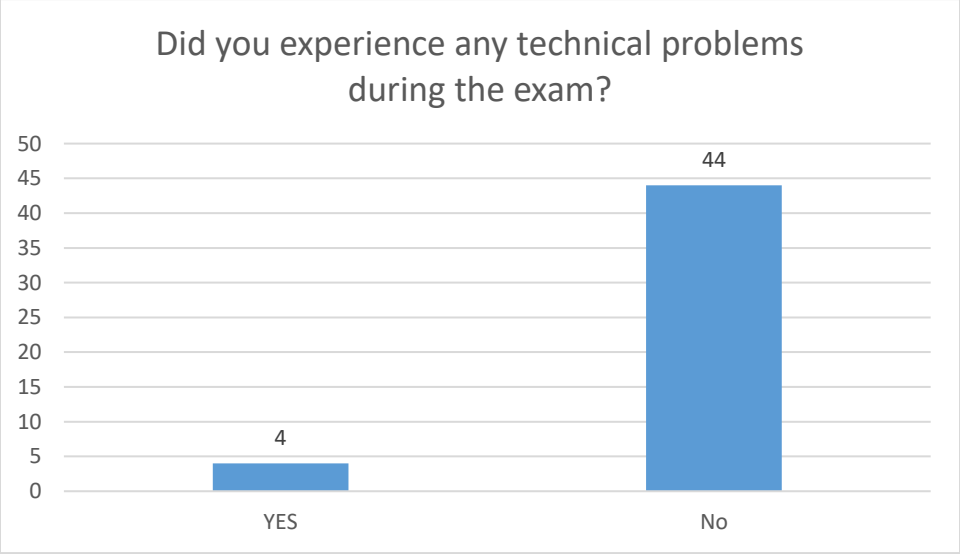


Figure 6.18: Technical problems experienced during the examination

The following section is the qualitative analysis of the focus group discussion followed by the validity framework analysis.

6.3 Focus group results

A total of 33 EM residents participated in the focus group from three regions. Twenty-six were male and seven were female. All groups were mixed gender except for one group. Two of the focus group conducted in the central region were residents who were from the MM group only. One of them consisted of junior residents the other of seniors. The rest of the three focus groups were of mixed levels of residents and both text and multimedia groups. The reason for this was to have a dynamic discussion where perspectives of both views were captured regarding the test items. Table 6.39 represents the demographic information of the participants.

Table 6.39: Demographic data for focus group participants

Focus Group	Region	Format	Level	Number	Gender (M:F)
One	Central (Riyadh)	MM	Juniors	4	3:1
Two	Central (Riyadh)	MM-TXT	Mixed	5	male
Three	Central (Riyadh)	MM	Seniors	7	6:1
Four	Western(Jeddah)	MM-TXT	Mixed	8	6:2
Five	Eastern (Khobar)	MM-TXT	Mixed	9	6:3

A pair of 18 items were displayed (in total 36 items) once in the multimedia format and another in the text format. Items were displayed and residents were asked to read the question, answer it, and explain why they selected the answer they did. They were also asked to identify any factors that helped or hinder their decision-making process to identify the answers. After each item, a description of the item analysis was displayed and reasons for why items behaved as they did was postulated and explained by the residents from their point of view. Appendix 21 presents an example of one of the items

in a combined view of item analysis of an item with comments made on the item from residents. It also presents the item in both formats MM and TXT, as well as the relevant theme for each quote.

6.3.1 Transcription themes and codes

Thematic analysis of the transcription using Braun and Clark's (2006) six-phase guide revealed seven main themes that emerged from the discussion, each containing a group of subthemes. A total of 1114 final statements were extracted and considered. The main themes are demonstrated in Figure 6.19 and were related to the clarity of the questions, quality of the multimedia, issues related to computer-based testing, characteristics of multimedia, issues related to time, level of question difficulty, and miscellaneous factors.

Theme	Sub-Theme
1.Clarity of Question (30%)	Content Clarity of Content Clues and Cues Language & Wording Missing, Incomplete or unnecessary Information Position of Information Incorrect or More than one answer Item Format Presence/Absence of Multimedia Expectation from Examiner
2. Multimedia Quality (20%)	Clarity Orientation View & Labelling More than one condition/factor Severity of Condition Length and Size of multimedia Measurement Tool
3. CBT Exam Setting (13%)	Break (Refreshments) Environment (Seating, Venue, Registration, Supervision) Computer issues (Functions, Technical) Proportion and number of items
4. Characteristics of Multimedia (11%)	Acts as Supplementary Material Clinically Oriented Provides unanticipated information Relays Enough information (confirm/spot diagnosis) Stimulate Cognitive Function Visually Oriented
5. Time (10%)	Enough Time Importance of Time (save, waste, analyse)
6. Question Difficulty Level (5%)	Item Complexity Knowledge level and skill Order of Item Residency Level
7. Miscellaneous (11%)	Personality Multimedia Preference Ethical issues Item Bias

Figure 6.19: Maine themes from the focus group discussion

6.3.1.1 Clarity of the question

The first theme, “clarity of the question” represents 30% of the statements extracted and describes factors that affected the clarity of the items. This theme had four subthemes that affected residents perceiving the items when read. Clarity of the items was affected mainly by its content, as well as other factors such as item format, presence or absence of multimedia, and expectations of items written by examiners.

6.3.1.1.1 Content of the item

Content of the item, which mainly affected item clarity, included factors such as content clarity, the presence of clues and cues in the item, wording and language used in the item, missing, incomplete or unnecessary information that was present in the item, position of the information in the item (whether it was presented in the stem or the options), and the presence of incorrect, protocol-driven or more than one answer. The following sections describe theses subthemes.

6.3.1.1.1.1 Clarity

Clarity of the item content as described by the residents related to issues that either made the question more or less clear to them. Sometimes, the image made the item clearer because it displayed enough description that was needed in comparison to text “*from the image we knew it’s mastoiditis if it’s only the question by itself we would say* its otitis media”. While other times, the image aided in excluding diagnosis “oesophageal rupture cannot be ruled out… *in the* X-ray the absence of mediastinal air *and these things I mean* kind of goes against, *I mean* it gives you a more wider range of exclusion of other diagnoses rather than description” as opposed to the text which did not “*because it can*

all come, *I mean* I can exclude *for instance* retropharyngeal *but* peri-tonsillar, *I can't, and* angioedema *I won't be able to*, *and* epiglottitis *I won't be able to".*

In some instances, the image acted the opposite way, making the item unclear while the description (text) was perceived as the clearer format "*especially the* ultrasound *of the* pregnant lady, *a lot didn't get this one. They got it wrong*, *but for us, it's there so we got it right"*. Some residents found that the content in some items had no relation to the image, which confused them and made the item unclear to answer, preferring to have had the text as the clearer format "*I know but the* history is not correlated *with the* image" *and "*child abuse, *the* history *is different and the picture is different".* Even if the resident knew the answer the presence of the image didn't help "*Either way the* diagnosis *is not clear*, *right the answer is clear, but the* diagnosis *is not clear".*

6.3.1.1.1.2 Cues and clues

Cues and clues played roles in both formats of items and were what helped residents reach the diagnosis faster in some items or aided in answering the items immediately without continuing to read the questions. As one resident put it, "*The* text *because it will* give me exactly what I am looking for *and* that's it". In some items, the multimedia acted as the cue to reach the answer and, in other times, the text gave away the answer. For example, when multimedia acted as the cue because it was so clear, residents felt they didn't have to continue gathering information from the item: "*Yes, this is it without reading the* stem. *I will answer" and "I mean there were a lot who said from the picture we knew* the answer, *anyway I don't need to continue*, *I mean they didn't need to read the question".* In the text format, certain terms and descriptions led residents to either reach to an immediate diagnosis or translate the word meaning into a diagnosis "hyperdense

means bleed" or "*the last line gave you the answer*, goes over trigeminal nerve *that's it,* distribution". One resident pointed out the difference between both formats, where the text required no thinking and the multimedia required using visual skills "Hyperdense *means* acute stroke *that's* it *but this one (means MM) it's not* hyperdense. *I have to see it".* Another remarked had he gotten the text format he would have answered it easily: *"If I got* written vesicle, *as soon as you say that it's zoster".*

6.3.1.1.1.3 Language and wording

There were some instances where residents commented on words that were either not understandable, unfamiliar, or the choice of wording was not appropriate for the scenario and led to uncertainty and unclarity of the item, such as "There is soot; *I didn't understand* soot *anyways".* One resident pointed out that the way an item was written made the question of a more difficult level than it should be: "*The information that this question is testing is very easy, it's not difficult*, *but* the way it was written makes it a moderate difficult question". Another resident pointed out that the medical terminology that was used was incorrect: "*Anyways,* it's wrong to label it the gestational sac. Extrauterine adnexal mass *or* mass. *You don't say* gestational sac, *because you remember* probable, *the table of* Rosen, probable *and* possible *and…"*

6.3.1.1.1.4 Missing, incomplete or unnecessary information

This subtheme of content clarity covered statements related to items being unclear because of missing information that was felt needed, or added information that was felt unnecessary. Some residents felt that because the question was missing something it was unclear. "*And the question was unclear*, *there was no* video *or any* ECG, for instance,

to show you that". While others felt that the descriptions of some reports were incomplete compared to the actual visual report they would have seen in the ED and would help in the diagnosis: "*He didn't write that* aVR *is higher than the* V1, which is one hint to let you know *if there is* left main versus *the* LAD". Another resident said that if a test was added, it would have made the item clearer to answer: "*So I think* an extra information *in the* question that guides you *that* it's actually candida *would solve the problem".* In some cases, the items had more than enough information and were felt unnecessary to be included, as it was just distracting and made residents interpret information that was not required in the end as one resident said:

> *"The question, in the end, wants an investigation you don't need to list me the vitals and list me* extra information. *You made me read the paper and I don't care about it in the question."*

6.3.1.1.1.5 Position of the information

Sometimes, the presence of information early on in the stem or later in the options made the item clearer. The information, when present directly in the stem, helped to reach the diagnosis from the start with no misdirection. For example: "*This is* foreign body *it's from the* history". It could also be in the stem and felt to be unnecessary, but the combination of multimedia and options is what makes the question easy to answer. "I don't need the stem; *he gives me the picture*, *and gives me the* options. *The picture really* was *enough*; *I mean the one who chose the picture is an artist".* The information in the stem sometimes misdirected the thinking process of the residents and the options redirected them back to the correct answer:

V-1 “*No* abdominal *because*, *in the options he put* transvaginal, *so I mean this is* abdominal

V-2 “Ah so I have to read *the options so I know what is the picture*, *ok why don't you tell me that this is an* abdominal *question from the beginning, so I can think in a* structured *way better than this”*

6.3.1.1.1.6 Incorrect, more than one answer or protocol-driven answers

Residents debated on the correct answers in some items because items were felt to have more than one answer for the given questions. This was either due to a) mismatching of information in the question with the multimedia “*No, there are questions that are wrong*, *for instance,* eye discharge yellow, *and the answer is* chlamydia, *and it is supposed to be the* yellow gonorrhoea. *I put it as soon as I saw the picture”;* b) differences in hospital practice and management protocols among residents in training hospitals: “*Even here there is a* third possible answer *also*, official ultrasound, *I mean we have it in the hospital.* Still we *the* bedside ultrasound is not of official I mean to be medicolegal documentation”; c) answering based on what is done in practice or what is standard in textbook: “Practically, real-life we will get an official ultrasound, *but we know that in the* exam *they want us to put* obstetric consultation”; d) exposure to the case “*so all these answers can go with it*, *so if you didn't see a similar* case *to its situation*, you will not answer it”; e) debate on which treatment came first as explained by one resident:

> “I could understand why that could be written *because sometimes a lot of times* when you place this as an option it's considered resuscitation option *ok. For instance*, *you find* A. is put as resuscitation, B, C and D *are put as* procedures, *so* you’re torn between resuscitation then procedure, or procedure then resuscitation.”

And; f) related to a certain speciality “*But wait a minute there is something important*. *It's known in* derma *you sit with two* consultants *in the same clinic*, *each one says the* diagnosis *is different from the other one”.* All these mentioned points were related to the clarity of content, which was a subtheme for “Clarity of the Question”. The next three sub-themes also affect question clarity.

6.3.1.1.2 Item format

Item format had an effect on how the content was perceived by the residents as either more or less clear. It also affected the way an item was interpreted. Item format also seemed to affect the level of cognition used by the residents and way of thinking. One resident explained how the thought process was in a CT scan and text item (Appendix 21):

> *“The idea is, the* image first you will see, interpret the image, after that you will see there is early sign of ischemic stroke, then you will think I should give tPA but there is contraindication by the vital, the blood pressure, so I should decrease the blood pressure before that. Then I will give the tPA, so first choice will be to give labetalol which will reduce the blood pressure. While in the other question in the written one, you will see the right hyperdense middle cerebral artery, its infarction, you will go to the vital directly, it's high, give labetalol directly.”

Regarding cognition level, some residents viewed multimedia as possessing a recall level of cognition “*No*, you are recalling an image of pneumothorax that you know already, you know that this is pneumothorax *right*, *the* radiolucent area? *but someone* describes it for you, you are seeing an X-ray, showing radiolucent area in all the border of the chest. *With this (text), this makes you* imagine more and use more thinking, maybe it will let you miss the diagnosis”. Another resident felt that multimedia made them think of more differentials: “*When we look at the picture, we would think of more than one* diagnosis *I mean really, to be honest, this is very important those that got the text they answered it correctly”.*

Depending on the format, residents would reach different diagnoses for some items: “V-BD1: nasal foreign body by text *by the* picture *I don't know” and “*V-NP: *yes the one on the right (text) I will choose* beta HCG *while I'm smiling*, *but the one on the left (MM)* I will not do beta-HCG regardless what the best result say”. Text formats were straightforward and required no thinking process for some residents “V-5: *for example, what answered for me is that I don't need to think* free fluid, no fluid, *is there foetal* monitoring, so they gave it to me” and, for some, the text format provided part of or the whole answer: “V-K: *In the* written part *he gave me half of the answer, in the* other (MM) *no he didn't give me half of the answer*, this is an extra level” “V-2: *You made it obvious for him the answer*, you gave them the answer”.

6.3.1.1.3 Presence or absence of the image

In some cases, the presence of the multimedia made the question more confusing. In some items, residents felt that the image was not needed and only added confusion to the item or wasted their time in looking at it: “*It has no use. What will it add? Just a picture of someone*, and *on the contrary, it will confuse me, do you know what confuses me in it. See, there is even a rash on his face and it is mixing up the topic now”* and “*notice that we know* Ramzi Hunt *and we know* Herpes zoster *and we don't know the answer because of the* video”. One resident explained his point of view regarding adding images:

> “*I don't know how was the arrangements of the questions I mean some of the* questions it didn't need *at all* MM *and the MM was put in it. Like the question of the* tamponade, *because the stem was long and above that, there was a video of a tamponade and when we went back to the questions he had only wanted (*what's the next step?). *So, if we are* simply saying that this is a cardiac tamponade I would have at least saved time have for the thinking process *because this is the most questions that we got mixed up with* either because *there was no* time in regards *to what it is, not to mention the options it wasn't* clear”

While other residents felt that the absence of multimedia was not good, and felt that the item needed an image or video to make the question clearer to be able to answer. Or the image needed extra information to complete the scenario "*Even here, I expect here, there has to be an image*. *It will make a difference".* One resident described what a good picture resembled: "*So* if we're going to talk about clinically, *I mean* clinically appropriate question to give a picture, then it needs to be a picture that's going to add. *I mean* it didn't. This didn't add much *because they* already *wrote that he has* white patches *and* change the colour of the mouth. What will add *is to bring me a picture* of the scrapable or not, a diagnostic method."

6.3.1.1.4 Expectations from examiners (overthinking)

This theme has to do with pre-conceptions from residents and their expectations regarding what the items should include and what the examiner (i.e., item write) expects from them. Some senior residents expected that when a clinical manifestation was presented it should be very clear (exaggerated) or only of the typical classical presentation. If not presented this way then it drove their thinking to another direction or made them overthink as they did not expect it to come in the exam. One resident said:

> *"Yes, but what is the point? that the* test is different b*ecause it's an* artificial environment. *I have to* over exaggerate *the* clinical sign *so that they can see it I mean the* candidates. *Maybe this* in real life they will make it a good diagnosis the right diagnosis, *but for me, in the* exam I'm expecting the over-exaggeration *so if it is* sudden *like this*, *I don't expect the* examiner *to bring me something* sudden like this. *They will bring me something very clear."*

Another resident also said:

> "*But I didn't think in the* exam *he would bring me something that isn't* classical…when they bring it a presentation, it has to be the classical. *It's supposed to be. They don't bring* something out of the normal *right*? *Bringing a picture* with something that is not that classical presentation drives your thinking into a completely different pathway"

Some residents tended to overthink the item and the multimedia that was included in the item, which might make them miss something: "*I'm* analyzing *the* image more, *but at the same time*, *it increases the* risk *that* I miss something" or made them select the incorrect answer:

> "I remember my thinking process at the time *that* I was thinking ok so they give me the multimedia because they want you to do the procedure, not to resuscitate, although the correct action would be to start resuscitation if the patient does not respond. Then you do the pericardium centesis. *So that's why I felt the* image might have confused my thinking process."

6.3.1.2 Multimedia quality

The second theme that accounted for 20% of the statements made by the residents had to do with the quality of the multimedia itself. Analysing the content in this theme led to six sub-themes that affected multimedia, which, in turn, affected the residents' performance towards the items. These were clarity of the multimedia, the need for labelling and orientation to the multimedia, the presence of more than one condition in a single media, the size and length of the multimedia, the severity of the condition displayed and the need for a measurement tool of some sort accompanied with the multimedia.

6.3.1.2.1 Clarity of multimedia

Most references to multimedia quality were due to clarity of the multimedia used. This had to do with the selected image or clip that didn't appear to relay enough or appropriate information to answer the question. This was expressed by one resident through the need to repeat the video to try to answer the question: "*So the* video *was that one, I repeated it without exaggeration five times and in the end, I answered and I wasn't sure".* Clarity of the media was expressed to be given in the exam setting very clearly in order to be able to answer the question: *"*Heartbeat *you have to either bring the picture clear like how you see it in* real life*. Anyway, even in* real life sometimes you're asking your colleagues for second eye to see there is no heartbeat, to confirm that". Another resident said: "I agree, *not always* through our clinical practice *the picture becomes clear like google, but* at least for exam wise to know the sign, *and* anything else I can manage". Clarity of the image might play a role in how the image is perceived by the examinee: "In general, this might look a little bit elevated to me *but* does not look elevated to *N*, *so here comes the part of bringing a clear* image *that no two would differ on that this is* ST elevation"*.* Some residents gave conflicting views regarding the clarity of the same images; where some could see it clearly and others could not. Other residents, when viewing both formats of the item, felt that the image was not clear compared to the text: "*Yes, and also if you've noticed*, *the* ear *here was not clear, in the description it wrote it* clearly"*.* And another commented on the choice of media included as being bad: *"I mean the problem is not the presence or absence of the image*, *the image itself is bad".* Some residents felt that the multimedia used were acceptable and clear and attributed clarity to computer issues: *"Maybe the screen itself didn't serve a lot, but* it was acceptable". Some residents felt that

multimedia items were clearer and easier to answer: “*But what was its percentage*, I think 80% *or* 85 % *was clear*, *and that you are discussing on* three, four.”

6.3.1.2.2 Orientation, view, and labelling

This was an important theme in multimedia quality and had to do with the spatial orientation of the multimedia, the organ, and space around the condition being explained. In addition, the cut, view of the image presented, marking and labelling directions were mentioned. These were all felt to be highly important to residents when interpreting multimedia. Viewing the image properly had to do with a) being able to manoeuvre the media in order to get more than one view and have a better understanding of what was being seen: “You have to take time; you have to manipulate; you have to take two views (transverse and longitudinal). *Orientation was important to be able to understand what part of the body one was looking at. One resident explained:*

> “*Before I read the question, I thought it was* abdominal aorta. *When I read the question, I said ok this is something in the* pelvis *so* I start re-thinking. *Of course, this takes from me a lot of time*. I try to find out *what exactly is there here*. Is that intrauterine or this is *out? Where is the* probe *exactly*, where is he directing the probe?”

Labelling was viewed by residents as important to have a better understanding of what they were looking at. This would save time in orienting themselves to the image. One resident even commented that it would be fairer to the juniors “*If this image, for instance, had an* area, they mark the area, *maybe it would be* more fair *for* juniors”. Sometimes the image was too close or the video too short and confused the resident: “*Wait a sec but I mean* it took me a while to realize what we are looking *because* it's too zoomed in. *In my mind, I thought that* it's an omphalocele or something. *Why don't they write for us*

labelling? *I don't know coronal* or sagittal view of the pelvis *for example*, *because* if you are not an expert *in the* ultrasound you might not realize *what the* clip is. *It's* too short for you to realize in a way".

Residents commented on that they wanted to be able to manipulate the image brightness, size, and cut too as done in the clinical setting to be able to have a more complete view of the lesion or condition: "I'm talking about images as X-rays *and like that*, *the problem is* sometimes it need *that you have to adjust the* brightness, *like the* CT *for example*, *you have to adjust the* brightness, *you have to adjust.* You see it from a different angle from different sizes *and like that".* Another said: "And zoom out zoom in. *You play with the resolution make it high, low".*

6.3.1.2.3 More than one condition or factor

This implies to the multimedia not purely containing one condition in the clinical presentation. For example, one of the cases was a child presenting with a foreign body in the nose, but also had an area of acne and redness on his face. Although the redness had no relation to the scenario it did distract some residents from answering the item. "See *you can say it is* allergic rhinitis *from the options because there is some* rash". Another case was a patient that had a skin manifestation of zoster that was on the trigeminal nerve distribution but also had crusting, which was another infection superimposed on it. This confused some residents of all levels. "This is very misleading. *There is* crust, honey crust appearance, *and* superinfection, *and maybe she has* superinfection *in reality*, *but he wants you to answer the* herpes zoster, *with the picture like that*. *If you brought another* dermatome *for instance*." One resident compared both formats saying:

> "*The* distribution *is with the* nerve herpes zoster, *but I mean when you see the question,* the text *will answer* herpes zoster *without thinking. But when I see the picture I will get mixed up. I will say maybe he meant the superinfection of the impetigo, or"*

6.3.1.2.4 Severity of the condition

This has to do with how severe the condition appeared to the residents perceived from the picture to what they know about the condition. Some residents would view the multimedia and feel that the condition is benign or not severe enough to what the item writer intended. This was compared to the description in the text format and in some items, it demonstrated that the image did not fit the description intended for the scenario. "B*ut the text doesn't* give you an idea of how sever is … *he didn't show you the* severity *of the* image" or "*It looks* more benign".

6.3.1.2.5 Length and size of multimedia

This was more related to the technical aspect of the multimedia; the length and duration of the clip that was used in the video, as well as the size of the image or video that was being viewed. Residents felt that some video clips were too short or too quick to be able to grasp and understand what was going on in the presentation and, therefore, didn't allow them to have a proper view from the first play. "*But it was about 20 seconds* or 30 *seconds, the video It was something like that"* or saying that the image was too fast *"See how it is transforming* between *the* layers *very fast".* One senior resident expressed the importance of labelling the ultrasound because the clip was too short to identify the organs seen.

Residents expressed their need to review and repeat the videos because the duration was too short for them to diagnose the clinical presentation presented: "*By the time you do this it's over*, *you go back and repeat and that's it".* And one resident felt that the looping of the video distracted him: "The duration of the videos *I had a problem with it*, *some of them were two seconds or three the* ultrasound*. It is going for three seconds and* you know it or you don't know it *and it is repeating, repeating, repeating".* While another resident felt that the looping was fine and not distracting: "*If it was to do with the* looping, *I think like this is ok".*

The size of the multimedia also played a role in the clarity of the questions. Residents felt that some of the MM was not large enough for their viewing and they had to concentrate, repeat or take longer to view the media: "*It was* short *yes and* for me *to be honest, it wasn't big*, *the picture was, I mean I can enlarge it but the picture itself for the* ultrasound. *That's to say I had to highly concentrate"*. Even though the function of enlargement was enabled in the exam, some residents felt that some images were still not clear. This was also reflected in the survey results when some residents felt that the use of enlargement was not useful (Figure 6.14). One resident commented: "*If you enlarge it, it becomes not that clear, yes it was some of them".* And another one commented on the quality of enlargement: "W*hen you do* zoom *on the* image, *the* quality is very, very bad". One comment was that "The images were too zoomed when enlarged and too far when in the normal window". Other residents preferred to keep the image small as it seemed to be clearer that way than when it was enlarged: "*I used to return it back small and I would try to get closer* and get a better look at it". For some, the enlarge function seemed to enlarge the frame of the media but not the image or video itself: "Y*es but it didn't enlarge. The*

video didn't enlarge." Other residents didn't comment and felt that the multimedia was good enough as originally displayed even though it didn't enlarge*.* "*To be honest* I didn't try to zoom it *more than how it was put".*

6.3.1.2.6 Measurement tool

The last sub-theme in MM quality relates to comments regarding comparisons to the normal clinical sign or the presence of a ruler marks that would help measure certain abnormalities. Residents felt that including these within the X-ray image as done in practice would have helped them answer and made the diagnosis more obvious: "If you give measurement *maybe I'll know"***.** In addition, residents felt that the ECG paper should be clearly viewed when enlarged to be able to count the small square units on the ECG graph. This is to measure the height (amplitude), as well as the width (breadth) of the waves. "In practice *when we have a doubt, we bring the paper and see* one square, two squares. *But on the screen, we can't do that"***.** When asked, some residents gave the opposite opinions and felt that the image was clear enough to answer and there was no need for any measurement tool to be included.

6.3.1.3 Issues related to CBT administration

The third theme that seemed to be emerging from the discussions were issues related to the CBT. This accounted for 13% of their statements that were extracted. Issues with CBT were to do with breaks, computer issues, environment, and the exam process.

6.3.1.3.1 Breaks

Most residents commented on the importance of having a protected break time during their three-hour examination. Breaks were needed for using the bathroom, as well as for having a rest from the examination. Residents felt anxious and unfair that there was no way to pause the exam time to go and use the bathroom. Bathroom breaks were necessary during examinations because of stress and the amount of coffee they had consumed early in the morning. Comments were: "*All of us need to go to the* toilet. *I mean I'm sorry* stress *we have to go"; "I needed to go to the bathroom, but I discovered that the time was running. I mean* this is unfair." And "*The* break from my own time, *I mean* I go to the exam after a few cups of coffee, I have to take a break". In addition, residents expressed that they needed a break to do whatever they wanted during the examination period, to have their own space, have refreshment, and relax a bit. *"Give me* space, 15 minutes *I can go and* drink water, *I breathe a bit,* stretch." Residents also commented on the issue of refreshments and having the need to drink or snack on something. Because the examination was in computer labs, residents were not allowed to have food or drinks inside. This bothered them as some felt thirsty and others felt stressed and wanted to have a drink of water "*I wanted to moisten my throat because I was* stressed".

6.3.1.3.2 Environment

Regarding the environment, residents commented on the seating, the heat of the room, and venue. One resident had felt that the room was hot while others felt that it was cold and some felt it was good. This was reflected in the questionnaire (Figure 6.17) where there were opinions of disagreements of the room temperature being comfortable and

were expressed in the focus group as stated: "*The hall was very hot, I was getting vagal in the exam from the heat"* or "*The room was a bit cold*". Residents were happy with the noise-cancelling headsets that made them concentrate and not hear background noises "*The nicest thing is the headset*, *for me this was the best thing"*. Another resident felt that even though the headset was available, he was easily distracted: "*Even though, no, coughing, I was distracted"*. One resident didn't feel comfortable using the head seat because of his glasses, others did not use it and some pointed out that this was not available to them in the Eastern Region.

Regarding seating tables, residents felt that they had their privacy during examination, and females felt comfortable sitting in the same room as their male colleagues: "*Anyways you don't even know who is surrounding you" and for females "If the* female *wants* privacy *she wants to take off her* niqab (face cover) *or something its ok".* The idea that all residents were present in one room brought comfort to them. They were sure that they all received the same treatment, supervision and information as commented by them: *"Leave us all in* the same area, *the same information will be said, the same* description"*.* Regarding registration, some residents felt that the registration process was complicated and made them feel uneasy before the exam. They felt that they just wanted to go in and start the test as one resident put it:

> "*The way of* registration *it was I don't know that you come and take the ID card and register and then go back and sit in the rest area and then come back again. I mean it's a bit* complicated, *it's not* easy just come in *and that's it. I mean the idea is that you are coming you want to go into the exam and that's it. that's.. its* stress under stress *and like that, I want to go into the exam I don't want to sit another place and get distracted like that. I mean one is already psychologically* doesn't *want to".*

Regarding supervision, only one resident commented on the proctors being a source of

distraction during the examination: “*Those who were supervising us were distracting us honestly; I mean the going and coming”.*

6.3.1.3.3 Computer issues

There were both positive and negative comments towards the use of computers as a mode of exam conduction. Most residents felt comfortable and enjoyed the features and functions of the computer examination. This was also reflected in the questionnaire (Figure 6.16). Residents liked the feature of marking items that needed reviewing. This was also favoured in the questionnaire (Figure 6.16). But residents did comment that they were not able to use it fully due to time constraints: “*I mean there was a nice feature which was* re-mark, *and I will go back to it*, *but* you didn't give me any time”. In addition, the feature of highlight helped them highlight key features in the scenario and helped one resident to highlight what was important in long scenarios: “*I* highlight *because it's long*. S*o,* I highlight *the* symptoms *for instance or the* age *the things that I feel made a difference to the* answer” and because they were used to it in paper examination as a test strategy: “*Because*, usually *in* paper I underline *the* things that I don't want to miss, I know *that this is a* clue that I don't want to forget about so I used to highlight it. *There are things he writes on the basis* to deviate you from this *that's it* to rule out certain diagnosis”.

There was one feature where most residents expressed on the survey as being distracting, which was the scrolling function (Figure 6.16). However, during the focus group, only a few commented on it saying that it was good and wasn’t an issue. Of the comments regarding scrolling were: *“It was clear from below, it was written* scroll down I go down” or “You don't want to miss the image, *this is the* concern *that* some people might not realize that there is a scroll down option”. However, not all residents shared the same

view about the exam features. Some felt uncomfortable using them, or felt they were too complicated or there was no time to spare using them during the examination as expressed “In *the computer,* highlight *it takes a bit of time”;* “*We all wanted to,* we just want to finish the exam *and that's it nobody wants to think of trying something else”,* and “it was complicated”. Most residents had an exam with no technical problems, with only a few complaining about technical issues they had faced during their examination. These are probably the few that have answered yes to the technical question in the questionnaire (Figure 6.18). A few residents had their video play automatically without them starting it and it confused them when reading the question and distracted them from how they wanted to approach the item. One resident explained:

> “*When I put* next, *the* video automatically starts, so it was distracting *that I read the question so I start to be in a daze, I mean* I'm assuming. *I start guessing what is the question before I even read the question*”

Another resident expressed the same view saying: “*The video works automatically*, so to me, to be honest, I would get preoccupied. *I don't read the questions*. I'm assuming *what is the question from the picture.”*

6.3.1.3.4 Proportion and Number of Items

Lastly, residents commented on the number of items in the exam that exhausted them and related to why they needed a break in their examinations. Residents seemed to agree that when they reached 50-70 questions, they felt tired and started to lose concentration: “*Especially in the exam situation, the* deflates *starts after 50 questions you can't concentrate well*”; “I mean I reach question *for instance* 60 *or* 70 *that's it. I start to, even if the question is easy, my concentration is not going to be as the beginning of the questions”.*

6.3.1.4 Characteristics of multimedia

This is the fourth theme that highlights the properties and characteristics of multimedia items as expressed by the residents. It comprised 11% of the statements and contained six sub-themes: clinically oriented, visually oriented, acts as supplementary material, provides unanticipated information, relays enough information, and stimulates cognitive thinking.

6.3.1.4.1 Clinically and visually oriented

Residents highly expressed that multimedia items were clinically oriented and they preferred these types of questions; as it reflected what they did in practice on a daily basis "*The* exam, in general, *was* clinical, *I mean honestly there isn't the system go read the book and come*, *no, it was clinical"* and "*I mean*, *I see it as if it's a* case *coming to the hospital*. *It's not just the memorizing what's in the book and you come*". A few residents commented that it makes for a good clinician in practice to know how to answer these questions. One resident said: "*This* makes you good physician, to judge your physician is good or not where you will graduate physician, have your patient safe". It also helped one resident to not miss information as in text format: "*The* text you cannot imagine by it. You cannot get justification for the *picture. But this* this is clinical. *I see it like this they come to you like this you* say its classical textbook; *but here* (text) you cannot, *sometimes you don't read the question properly, you're in a hurry* you don't register it from the history". Another resident pointed out that using multimedia goes more with clinical scenarios "It will be more in context with the clinical scenario".

Another characteristic that was noticed was that multimedia items were visually oriented and were based on what residents did in their practice "see cases and read reports".

Residents expressed that they were used to seeing the cases rather than remembering them some of what was said: "*I think that because our* speciality is visually oriented"; "*Maybe we are used to the* ECG picture" and "*Because* this is what we practice, this is what we see". One resident said that getting an ECG in the test took longer to answer because they were used to seeing it: "*When he gives us* text, *describes in it the* ECG, we need time to imagine *because most of the* ECG *is about* visual learning".

6.3.1.4.2 Acts as a supplementary material

Multimedia items, as noted by some residents, don't always relay all the information; instead, they act as a supplement or hint to the rest of the information in the items. One resident said: "*It doesn't become* spot diagnosis *for the picture itself*, *I mean there is* description *for the* case, *you don't depend a complete dependence on it. Why? because sometimes the picture* is not the only thing that you need."

6.3.1.4.3 Provides unanticipated information and relays enough information

It was noticed that multimedia items provided unanticipated or additional information that would otherwise be absent in the text format. However, in MM, it would aid in answering the question. This extra information would be related to the background of the image or surroundings that would normally not be described in the text version of the scenario probably due to its irrelevancy to the case. For example, one resident commented: "*Here the* ear *was covered with the hair".* The absence of signs in the image aided in excluding diagnosis that text couldn't do: "*So from the* history *you will know it's either* A *or* C. *The* X-ray *will differentiate it for you*. *If there is* thumbprint its epiglottitis, *no* thumbprint peritonsillitis."

Multimedia helped in directing thought process in one direction as one resident puts it: “It just direct you to one place *maybe I wasn't thinking about it but* it actually directs me to one way*”.* It also provided the course of a disease that might not be included in the description: “V-S: *The* picture *added that* the rash started *in the* lower leg *and* then spread upward *The idea is the* pattern *of the* rash where does it start where does it end and the way it looks. *Because in the picture it's true that it didn't say where it started and how, but the* description of the rash *was there, and it wasn't there in the text”.* MM also provided missing information that was not thought of when writing the item: “Keeping in mind *that this is from the* text *he didn't say* this is flexors *or* extensors”; “*I don't sit and imagine the* lead, *ok how much is the* elevation, minor elevation, high elevation.”

In addition, for providing unanticipated information, it also has the capability to provide enough information to answer the question or to make the diagnosis quickly when chosen properly. Some residents described it as recall: “*I mean if for example, he got a* video something you have to recall it, *but you know the* video *this shape*, *and you saw the same* video, we'll answer the question directly” and “You are recalling an image of pneumothorax that you know already”. Another resident said: “Y*es when you see the* X-ray, *it tells you someone comes to you with* shortness of breath”. Sometimes, the multimedia was so clear, it acted as a cue and the question didn’t need reading: “Anyways *I am not going to read the question”,* or an indicator *“This is* spot diagnosis *the* picture I mean”. It also gave clarification to doubts: “*It will resolve the debate”.*

6.3.1.4.4 Stimulate cognitive function

Residents commented that MM items challenged their thinking and made them think of more differentials: “Put more differential diagnosis” and “*So it makes you use this*; *I mean*

cognitive thinking". As expressed by one resident:

> "*Like for example,* rash *when he writes it as* description, *I will answer it* term *we know it and write it*. *But when you bring a picture of a rash*, *despite that we hate* rash ok, *I will sit and think is it this one or this one or this one*. This is going to test me as I know it or I don't know it"

One resident said that it makes him use analysis but may also make him miss information: "*I'm* analysing *the* image more, *but at the same time*, *it increases the* risk *that* I miss something". Other comments by residents were related to mental feelings: "There is clinical, I mean challenging"; "*It's amazing because* it tests something". Comments were also related that to answer the question, the basic thing one needs was knowledge: "I mean the image depend on who knows what is it. I mean knowledge" and interpretation skills "*The one who doesn't know how to read the* X-ray he will not guess it". One resident commented that it serves as a teaching method: "It's *what measures for us the* real level of us *and it's really beneficial for us and its' what teaches us"*

6.3.1.5 Time

From the survey response (Figure 6.12), it was apparent that time was an issue for residents as 27% of the residents said that time wasn't enough. Time is the fifth theme, and although it accounts for 10% of the statements, it was almost always the first thing mentioned by most residents in all focus groups and was strongly voiced as one resident said: "*Of course the time, everyone talked about it*, which is something, *I mean, has to* be addressed". Residents had said that time felt short and that although they had time to complete the exam, there was no time to review it. Time was expressed by residents in different ways. Some said that there wasn't enough time to review the items: "*Maybe sometimes you have* enough time to finish the exam *but maybe there are things you have,*

you put a mark on it you want to come back to it, you don't have enough time." Others said that there was just enough time to complete the exam: "*The last questions and the screen turned off"; "There was 20 seconds left".* Two residents felt rushed: "*I used to feel that I am in a race honestly with time"; "For me it was, click, click, click quickly let me finish I mean* catch whatever you can". One commented that even if there was time, they wouldn't have energy to review: "Even if I finish early, exhausted to review because mentally exhausted I just finish I can't review anymore". Some resident felt that time was good.

Time was also expressed as being wasted by irrelevant information or MM or being saved because of cues and clear MM. Wasted time was due to difficult items: "I know I will not answer it I don't want to waste time *and I go to the one after it";* irrelevant MM: "It takes time *from me for nothing";* preoccupied if the item was marked or not: *"Put for me the* next 10 questions it will be for study"; interpreting MM: "You *will waste time";* interpreting results: "*If he wrote all* normal *and he's wasting my time"*; reading long questions: "*The question look how long it is, and the* diagnosis is *from the picture"* and as mentioned, orienting to the MM.

Comments related to saving time were mostly related to MM being clearer than text: "*When I saw it, immediately* I diagnose it"; "*The time for me to imagine these two and a half lines the picture was faster" and "The picture is easier and even the time".* Others felt the opposite that text gave them more time: "*I mean, I get the* description *maybe I'll have five minutes left".* Extra time was felt needed for MM items mostly for two reasons: a) to repeat and look again at the MM: "*It's clear, but it took time for me to repeat"* or "*And repeat and look at is there* bleed *or no* bleed", and b) to familiarize, and analyse the MM:

“You need time to interpret them to answer”, “It takes time, *this is* chest *this is* abdomen" and “You are not holding the probe and you know the orientation, *for example in the* picture *or the..* this is the issue; *I mean* these questions needs more time”.

6.3.1.6 Difficulty of Item

Difficulty is the sixth theme and accounted for 5% of comments. Most difficulty comments were either because of different levels of training, where items would be more difficult for juniors than seniors: “*But* R2 and R3 we *are alike”*, “*Because of the* level” or blueprint content not expected for juniors: “*Do you think the* R1 *will read the* spiral fracture”. Knowledge background played a role in perceiving the item as easy or difficult: “*I'm telling you if it comes to me in R1, I will say this is milk, I don't know except* Candida*”.* With more knowledge, more differentials are available and the item will differ between levels: “With increasing knowledge you get more options in your list”. Items will also differ with experience, as one resident said: “*The* experience *refines the information that you have”.* The other factor that some commented on was the order of the items in terms of difficulty level. Some residents liked to start with easy questions then have the items escalate in difficulty as one described it: “It gives you *a bit of* shake to the mind” and it boosts confidence “*But when you put* the question you gain confidence”. Others preferred to have the difficult items first because, at the end of the exam, they lacked energy: “*Let it be the first thing* difficult”; “By the time you get the difficult, you don't have any more energy”. Some preferred having easy items at the end because of less time and energy as one commented: “*The last* 30 *questions I answered it may be in* 30 *minutes”.* One resident didn’t agree on having difficult items all at once as it will affect the rest of his performance; he put it: “*That's why I messed up the rest”.* A few comments related to item difficulty was

due to the complexity of the item, with some favouring this: “You want a level of sophistication, the picture”. The extra level of thinking between formats renders the level of the item as one resident explained”:

> “*In the* written part *he gave me half of the answer in the* other (MM) *no he didn't give me half of the answer*, this is an extra level”

6.3.1.7 Miscellaneous

The final theme that accounted for 11% was composed of various factors related to learning preference, personality characteristics, and test-taking strategies. Some residents perceived the item differently because of their style as one resident said: “*It depends on the person’s nature if he's an, auditory person or visual”.* Residents who were aware of their style knew which item format they preferred: *“I know myself I'm visual”* and “I lean toward the text”. A few commented that the item will be perceived differently according to the person’s personality as one said: “*It also depends on the* personality” and another commented *“*different personalities”.

Other residents noted that certain questions should be presented using MM, such as ultrasounds: “*I'm* still *with the* images, *except in* GYNE Ultrasounds”, echocardiography “ECHO-abnormal wall motion, *this you need a* video”, rashes: “Even rash, *it’s better to come as a picture”* and ECGs: “*I think any* ECG *question* has to be ECG”.

Regarding item bias, almost all comments were related to one of the items used regarding a transmitted disease. Some residents felt that the mention of a country was inappropriate and stereotyping it. Comments were “*My objection on it* is for stereotyping”; “Don't stigmatize *the* geographic”. However, these comments were mentioned by a few. The rest of the residents felt it important to know the geographic distribution and

socioeconomic status of certain diseases.

Lastly, regarding ethical comments, some residents voiced that they didn't like taking the unmarked items during their examinations and that the blueprint had changed. Some felt that they weren't informed and others felt that they received the information late. On elaborating with the residents, the problem was highlighted due to miscommunication, as there was a change in the EM admins responsible for informing residents, and the newly appointed admin did not relay information in a timely manner.

6.4 Validity and validity framework

In this section, the results of using the framework for the purpose of this research are covered. The aim of using the validity framework was to follow a systematic approach to conduct and collect information and evidence throughout the whole process of the research study and use these evidence as support or threat for data interpretation to be able to make a sound and valid judgment of the results (159). Here, the evidence of validity focuses on the trustworthiness of the decision that is made based on the judgments on scores and results collected from an assessment procedure in a specific context (37). When going through the process of validity, the aim is not to conclude that the test is valid. Rather, it is to state as certainly as possible to what degree of validity does the test presume to have (153). Analysis in this section is through a narrative report going step by step what seemed to work in the framework and what didn't. This involves providing examples of what was done throughout the process as providing evidence, as well as threats that were faced during this research and providing solutions and explanations when appropriate.

6.4.1 General overview of frameworks

In assessment, depending on the literature and framework one chooses to follow, there are always at least three links in the chain of inferences that are needed in order to be able to interpret a score (67):

- Evaluating (scoring) the observed performance or (construct demonstrated)—deciding if it's good, poor, or somewhat in between

- Generalization of results from observed performance (construct) to other tests that are similarly designed, but not identical.
- Extrapolation of exam results from the context of assessment (e.g., simulation) to expected performance in actual practice

To demonstrate the three points above, with this research Figure 6.20 depicts this complicated process. If the purpose of the test is to identify the examinee's ability to make appropriate decisions for patient care (using higher-cognitive skills) in the speciality of Emergency medicine in a written examination; therefore, MM -MCQs were constructed, written, and reviewed well for this purpose, according to well-developed test content, according to item writing instructions, with questions that contained stems describing clinical situations that required examinees to indicate the next step in patient care. For checkpoints 1, the evaluation link for score interpretation should be very strong and straightforward. Because of the relatively appropriate number of the MCQs, the scores are most likely to be reproducible (point 2) (i.e., they will relate to scores of similar examinations covering the same EM content with different items). Point 3 – extrapolation - usually tends to be the most difficult link and the weakest link of MCQ score interpretation. This is most plausible when the observed performance on the MCQs are very similar to the real performance in context (in the ED), which is not the case here, as performance patterns on written examinations (cognitive abilities) are not reflected in the performance pattern in the clinical situation (clinical abilities). In fact, it might be the opposite. Performing well on MCQs does not equal to performing well in the clinical setting. However, it can be said that if the test was constructed systematically well and contained items that are highly relevant and in a way that tests higher cognitive abilities,

then it may be plausible that poor-performing examinees (with low scores) will tend to not be able to provide safe practice and won't be able to demonstrate patient care in the actual setting. However, this cannot be said for examinees that got high scores. Observable high scores from examinees on a test do not translate that all of them will become safe practitioners (67, 228).

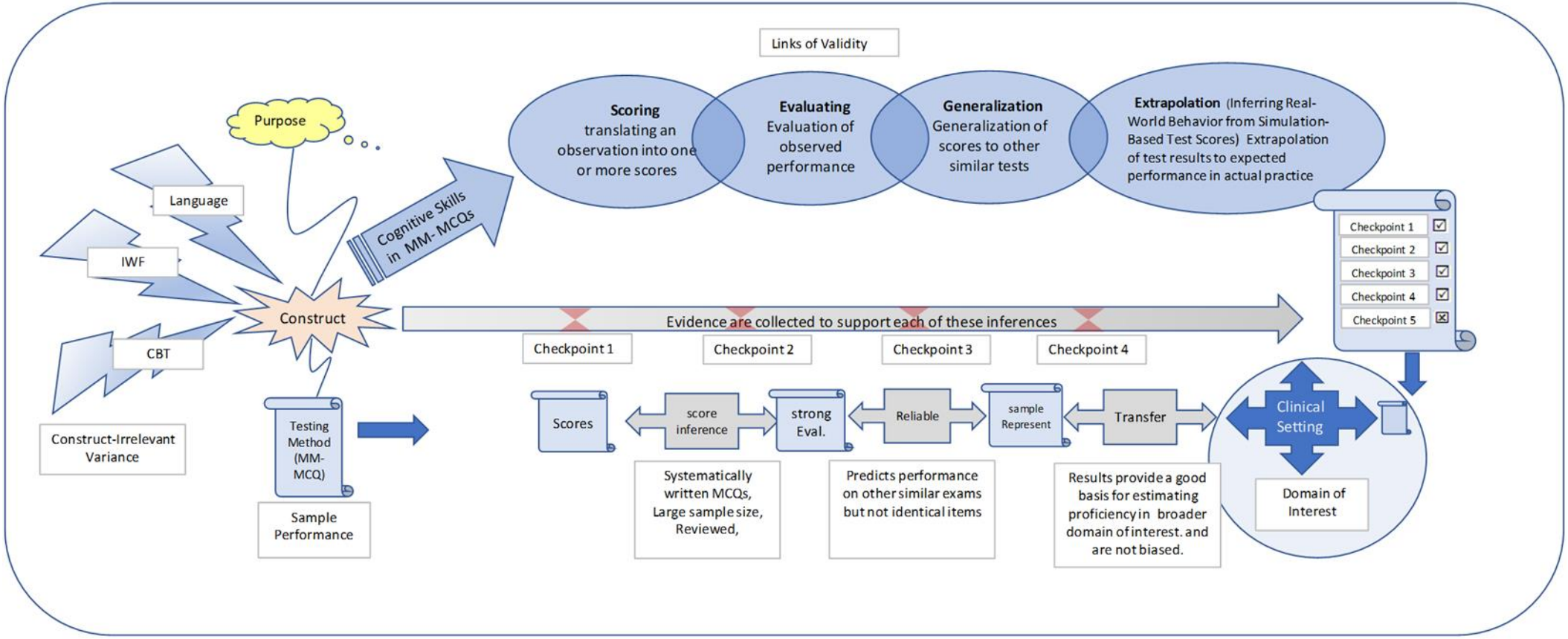


Figure 6.20: General overview of validity frameworks

6.4.2 Cambridge Framework

The framework was evaluated on how it demonstrated what is needed in a testing process (e.g., evidence, methods), who is expected to use it (i.e., who is responsible in each step), how to apply it successfully (i.e., clarity of classification and categories, ease of use, acceptability and value by users), and whether its applicability locally is as effective as its applicability internationally. The focus was on the first three inferences as the last two, as explained before, were not applicable in this research. Interpretations are also based on the proposed methods in the framework presented in Appendix 18. The framework conceptualises validity and helps determine the appropriate level and type of validity evidence needed for assessment. In many ways, it mirrors the framework of the APA (Downing SM 2003) that also has five headings; content representation, response process, internal structure, relationship to other variables, and consequences, as demonstrated in Appendix 7.

The display of the Cambridge framework in a table rather than text as other frameworks made it easier to comprehend and understand the concept. Identifying the i

interpretive and validity argument clarified what was described in Kane's framework. The presentation of questions and related methods made the framework more operational and accessible to apply. However, the proposed structure of the framework intended to be easy to understand and avoid technical language. This was not felt when dealing with the framework. The language and wording of the questions like most validity frameworks took quite some time to understand. And even till the end of the research, there was still some vagueness that seemed to be unclear. For example, the use of the word elicits in the questions made it vague and not easy to understand. Another example, in the warrant

section ‘scores reflect the quality of performances’ and its related question ‘are scores dependable measures of the intended construct?’ seem to be quite complex and heavy with information that is not relayed. The language and choices of words that carry heavy meanings would not be easy for someone outside this filed to comprehend. The researcher needed to explore other frameworks (Standard (155), Downing (164),(168)) for more details to understand the language used and compiled here. For novices in assessment and validity concepts, this framework is not easily understood. In fact, further reading and search of other frameworks and guidelines were needed in order to understand this one.

The framework claims that it can be used in high-stakes examinations. However, the display of the framework seemed to be confusing in some areas, particularly relating to content and the item development process as presented in Table 6.40. In the scoring inference methods again, the use of terms such as board documents and number of markers are uncommon terms in the researcher’s country (Saudi Arabia), as well as in the assessment literature. In addition, the first impression when reading scoring would be statistical analysis and item analysis in addition to reliability to be included in this section. It was confusing to see that the methods used in this section were only related to reliability of score results and methods of scoring, as well as classifying grades. Examples of statements are present in Table 6.40 The title of generalization inference, was misleading to the researcher, as the title implied like other frameworks the generalization from a test taker's observed score to an estimate of the test taker's universe score. However, in here it implies to the blueprint and if the task adequately reflects what is important in the syllabus. In addition, differences between methods in

generalization and construct representation seemed to be overlapping or unclear.

Table 6.40: Comments on Cambridge framework with examples

Inference	Comments and example
Construct Representation	Using general terms, such as examiner reports, in statements such as 'Review examiner reports for insights into how the questions were answered by candidates'. Does this indicate IA report or a report written by the examiner? In which case it's not applicable in MCQ examinations. Item analysis seems to focus mostly on IRT methods and gave the impression that if CTT was used that the validity evidence for the results was incorrect. Methods cover linking item to exam objective and cognitive level but no further reference here on the blueprint review, who and how should carry it out, who should write the items, and if any qualifications are needed. Reference to identify subtopics are mentioned in the generalization section
Scoring	Statement used in the framework 'Review exam board documents on marking and scoring procedures' does this refer to policies and procedures and if so, what are examples used. In addition, what does review imply? To make changes or to follow it even if not updated.
Generalization	The statement here was 'Ask appropriate examiners/experts to rate for each exam question the cognitive demands rewarded by each question, as reflected by the mark scheme'. Statement in construct representation is 'Ask appropriate examiners/experts to rate the extent to which each question places certain types of cognitive demands on students.' In the first statement is rating the cognition for each item based on IA? If so, is this for classifying the item? Or evaluating whether it fits with the rating of experts in the second statement?

Although covered by other frameworks and guidelines, this framework did not help in understanding the differences in test-taking strategies, behaviours, cultural references, linguistic references, or native expressions. These are all threats (alternate explanation) to validity arguments and are important to identify. The analysis of the research was mapped against the Cambridge framework and is presented in Table 6.41, and the following chapter discusses these results.

Table 6.41: Mapping research against the Cambridge framework for validity evidence

Points	Step in Cambridge Framework	Strengths	Weaknesses (Threats)	Measures Taken
Purpose of the test	Not specified clearly may be construct representation	Defined and identified using the K1, K2 taxonomy and considering the CLT.	Abstract concept of cognition that one can never be sure of	Follow guidelines and frameworks, do the best you can
Intended interpretation and use of test score	Not specified clearly may be construct representation	Intended interpretation was for test score use and was described	No other test formats were used, no criterion reference to relate to	One can compare results with other similar tests for convergence and dissimilar tests (e.g., OSCE) for divergence.
Item writer qualification	Not specified may be scoring, as a standardization method for the item writer	SMEs were content experts in the field guided by assessment experts during the whole process. SMEs were trained and selected	IW may have been distracted with some tasks or during Hajj and Ramadan	Items were reviewed more than once and by various members
Standard-setting	Scoring	Cut-score set, from a group of speciality experts, after item review. Cut scores are clearly documented.	Pre-determined fixed cut scores are the organisation's policy. Didn't change SS when including MM, didn't consider lower mean test scores due to MM.	Current SS that was applied here was norm referencing. This method is currently being revised to try and adopt criterion-referenced (Angoff) approach for future tests (55)
Scoring and score reports	Construct representation and Scoring	Every effort was made to capture test interpretation. Item statistics criteria and ranges were developed pre-examination. And post-examination was summarized and reviewed before reporting. Key validation was verified. Score reports were timely and distributed to the exam committee and	Scores might be affected by extraneous factors Candidates were not immediately informed of their results, post-hoc analysis by exam committee was completed before candidates were informed of their pass/fail status. In addition, performance on each	Categorised reports were later sent to program directors.

		announced to residents according to the organisations' pre-set rules	category was distributed later by training PD according to organisation policy.	
Misinterpretation of scores and negative consequences	Construct representation and Scoring	Results are first reviewed by Assessment experts with recommendations to EM experts.	Some analysis methods, such as IRT, DIF were not implemented.	CTT is documented well in the literature to be used. The addition of G-study and reliability was also implemented.
Time of exam is sufficient to complete the test	No reference in the framework for the exam process	Time was calculated to be 1 to 1.5 minutes per item based on work in the Middle East, for students whom English is not their native language. Also, based on the pilot study and following residents' requests an additional break was included in the exam. Time was thought of to view and answer the questions so there would be enough time to complete the entire exam (165).	Playback of videos, interpretation of more than one image in a single item wasn't factored in as additional time. So, although there was enough time to complete the exam, there was not enough time to review it. Some items may have been difficult for some junior residents and were considered a bit advanced for their level of training at their given stage. No additional time was factored in for MM items.	IA did not show a speeded test; almost all examinees completed the exam but didn't have time to review the exam from discussion and survey. A look into factoring in additional time for MM items will lead to the review of the cost of booking additional testing time with the publishing vender must be weighed against the economy (55).
Exam Setting	Not specified clearly may be construct representation	Scheduling the exam in a realistic formal setting (high-stake exam) adds to the generalizability of the results (54) Provides a realistic setting where candidates will perform on beta items the same way as to test items	Possible increased level of stress not knowing if the item is marked or not, and if there is enough time to answer	Orientation was given to residents explaining the new format. Increase testing time

Delivery Platform (CBT)	Not specified clearly may be construct representation	Has many advantages over conventional methods and can include innovative item formats. Improves construct representation Provides more IA information to judge on the quality of MM in comparison with the text	Unfamiliarity with CBT and computer functions may be a source of CIV Computer skills, complexity computer interphase, familiarity with navigating through its software, screen size, screen refresh rate may affect examinee performance, and technical problems (39, 47, 63, 67).	CBT more relaxing allows for presenting different item types, allows for in-depth analysis (e.g., time spent on item) (39). Guidelines for developing CBT were taken, and technical support and measure were available. Survey results demonstrated a good experience.
Instructions	Not specified clearly may be construct representation	Clear instructions were provided, presentation conducted. Provided orientation to residents and orientation before CBT, e-mails sent out with materials	Unclear instructions, not all residents understood it or forgot about it (67)	Results from the survey and focus group demonstrate that most residents understood instructions
Test specification (BP)	Construct representation	Developed and reviewed by content experts in EM field and medical education field. Content covered are important and relevant. The content reflected weight in the training curriculum Evidence for validity of the construct measured that senior residents should achieve higher scores (42).	The finalized blueprint was not distributed to residents except a couple of months before the exam due to the formation of a new exam committee that reviewed and updated the BP One exam for different levels may affect residents' perceptions towards the exam	Almost all residents were studying for the exam which contained the same BP, however, students changed their study methods Face validity (acceptability) of the exam was assessed by using a computerized questionnaire and Focus group feedback BP was checked for content and difficulty level of items to be

		Results showed that seniors have higher marks than juniors, expected because of knowledge.	Item pool created reflected the different grade level distribution, one would question if the sampling process of the exams produce fair and meaningful results	balanced according to level. Additional classification of the item was applied which was assigning a level to reflect that items were balanced.
Item Content	Construct representation	Items were in clinical vignette format developed by content expert and reviewed for technical and IWF.	The Use of items of low cognitive levels and violations of the item writing principles Irrelevant, difficult, not representative	The use of clinical vignette and MM materials to promote higher cognitive levels and clinically oriented cases. Developed according to reviewed BP.
IWF	Not specified clearly may be construct representation	Items were well-constructed and Items were reviewed for flaws: flaws related to testwiseness, flaws due to irrelevant difficulty, linguistic and Cultural bias (165) Review of items by more than one expert. Clarification from residents in the focus group.	Poorly constructed items can bias test with the presence of IWF (198) Linguistic aspects may affect examinees to understand the item Students for whom English is not their native language are at a disadvantage as their processing time is already slower than that of their native speaking peers. (198)	Issue of language seems to be overcome as it is introduced in school, and the 7 years of medical school are in English, by the time the student reaches residency he would have understood the language and has gotten used to the MCQ exams in the second language
Item Sensitivity and Fairness	Not specified clearly may be construct representation	Items are free of language and content that were offensive. Contains no elements extraneous to the construct, appropriate for level and do not contain emotional distress to examinee.	One item was felt to be inappropriate	Program directors and senior residents disagree and consider the item important to ensure residents know disease distribution

Multimedia	Not specified clearly may be construct representation	Straightforward, submitted and reviewed by experts. Most are of high quality and reviewed by external experts for clarity and fairness.	Some items may not have been of the highest quality (i.e., clearest) due to the degradation in quality that results from copying and reformatting videos, however, they were clear enough to answer the question. May accidentally cause CIV (39). Performance may be affected when irrelevant, excessive or additional information is presented e.g., irrelevant images (126).	Select high-quality images and videos, from their original source if possible. Otherwise, reduce the amount of copying and reformatting of such materials (19). MM are more authentic, clinically oriented and tests higher-cognitive skills (39)
Environment	Not specified clearly may be construct representation	Standardization of exam centres is equipped and checked by both SCFHS and testing agency to ensure that all testing centres are of the same standards. The environment was comfortable, secure, equipped with minimum distraction Three test centres help overcome bias of the overall results from a single test centre (109).	Probably, between the three SCFHS testing centres, factors such as the seating arrangements, temperature, lighting, noise, presence of other candidates in the testing room may have varied, thus affecting how well candidates performed on the test (19). Testing centre in one region had another speciality testing with them and did not have the headset for sound isolation Proctors may be a source of distraction to examinees during testing (189)	Majority of feedback from residents were mostly positive, with no major issues. IA showed no differences across regions. Standardization measures were taken for testing conditions and settings (165). Test sites took into consideration the comfort, space, privacy, lighting and environment that is free from distractions. Timing of test administration was comparable across sites. Proctors were trained and Security measures and procedures were taken to prevent deceitful behaviour.

Reliability	Generalization	Methods for selecting and synthesizing data from different sources (triangulation) and deciding when to stop (saturation) will inform the generalization inference for qualitative data (157). Increased number of items and testing time	Sampling techniques, sample size and non-randomization of participants weakens generalization to participants, context and situation that are not similar. (visual speciality, more than one level in one exam)	Reliability statistics: internal consistency, point biserial, G- and D-theory calculated.

The previous table demonstrated both the sources of validity evidence (strength) and possible sources of CIV (threats) that were faced during the test development process, as well as ways to overcome them. These are the columns of the second part of the Cambridge framework that represent the validity argument that was presented in Figure 5.7. To evaluate the validity of intended test score interpretations and uses and test psychometric quality, evidences must be explained and reported. Based on the qualitative data presented in Table 6.41 and the justification for each step, it can be concluded that the sources of the validity evidence gathered for this research are valid and interpretation made from the test results are trustworthy. To clarify the gathered sources, Table 6.42 generally outlines these sources against three frameworks: the 12-test development framework, the Cambridge framework, and Downing's framework based on the APA standards. The first column outlines the 12 components of an effective test development framework, the second column explains what evidence recommended by the framework was gathered during this research's test process, and the last two columns outline whether or not the Cambridge and Downing frameworks covered these issues in their proposed framework.

Table 6.42 Mapping sources of validity evidence gathered to the effective test development process, Cambridge and Downing's frameworks

Validity Framework (Sources of Validity Evidence)			
12 Component of Effective Test-Development Framework		**Cambridge FW**	**Downing/(APA)**
Test Development Components	Detailed plan for the test was explained	Was not mentioned or not clearly stated	Covered in the APA *Standards*
	Research phase was outlined and explained		
	Item writing process was explained		
	All test components, rationale and methods used were explained		
Domain Definition & Claims Statements	The domain and construct to be measured was defined and a clear statement of the purpose and claims of the MM-test were made	Construct representation	Test Content
	Test content and BP defined and reviewed	Construct representation Generalization	Test Content
	Examinee population defined	Not clearly stated	APA *Standards*
Content Specifications	Content specifications for item development was given	Construct representation	Test Content
	Form template, item format, ordering and section was provided	Not clearly stated	Test Content
	Test content, length and time allowed for testing was described	Construct representation	Test Content
	Score reporting, psychometric properties was shown	Construct representation	Response Process Internal Structure
	Directions for administrators, test takers and any other materials was clarified	Not stated	APA *Standards*
	Evidence from test results relating to the construct was explained	Generalization	Relationship to other variables
	CBT specifications (hardware and software requirements) were described	Not stated	APA *Standards*
Item Development	Appropriate item formats, materials were used	Construct representation and Scoring	Test Content Response Process Relations to other Variables.

	Items were written, reviewed and tested Process for item development, review and selection was illustrated and explained.	Not clearly stated	Test Content
	Expert judges training through workshops was given	Not clearly stated	APA *Standards*
	Procedure for pilot testing was conducted and explained	Not clearly stated	Test Content APA *Standards*
	Item model (CTT) and item analysis was used and explained.	Construct representation	Internal Structure
	Sources of irrelevant variance (language, cognitive, physical etc.) was demonstrated	Construct representation	Response Process
Test Design and Assembly	Test forms were designed and created based on test specifications	Scoring and Generalization	Test Content Response Process Internal Structure.
	Issues related to test content; format was attended to.	Not clearly stated	Test Content Internal Structure
	Procedures and steps taken were reported during test design, development and scoring to provide evidence of validity, reliability and fairness.	Scoring and Generalization	Internal Structure
Test Production	Procedures and steps taken were reported during test design, development and scoring to provide evidence of validity, reliability and fairness.	Scoring and Generalization	Test Content Validity & Consequence of testing
Test Administration	Test administration was conducted in a standardised way.	Not clearly stated	Test Content Validity & Consequence of testing
	Threats to validity that may arise during administration were identified and avoided.	Not clearly stated	Response Process
	Explanation, materials, sample questions and research purposes were prepared to test takers prior to their examination.	Not clearly stated	Response Process
	Documentation for test administration was explained to allow others to replicate the condition.	Not stated	APA *Standards*
Scoring	Policy and procedures for scoring, scoring criteria, quality control was followed.	Scoring	Response Process Internal Structure
Cut Scores	Cut score criteria, and rationale for procedures used was described and was consistent with the purpose of the test.	Scoring	Response Process
	Scores reflected the competency level of examinee	Construct Representation	Internal Structure Response Process

Test Score Reports	Test score reports that were accessible and understandable were provided to the examination committee and residents. Evidence of reliability and SEM was reported	Scoring and Generalization	Internal Structure Validity & Consequence of testing
	How scores were interpreted and used was presented	Construct Representation	Internal Structure
	Test security, score confidentiality and research purposes were reported.	Not stated	Response Process APA *Standards*
Test Security	Procedures for ensuring test security during test development and administration was described	Not clearly stated	APA *Standards*
	Handling of copyright tests, material transmission, score and test confidentiality and data for research purposes was maintained and reported.	Not clearly stated	APA *Standards*
Test Documentation	Procedures and steps taken during test design, development, scoring and analysis were provided.	All sources of the framework	All sources of the framework
	Population, sample characteristics, material used, language and translation procedures were reported.	Not clearly stated	APA *Standards*
	Procedures for the administration of computerized tests and tests using multimedia was explained.	Not clearly stated	APA *Standards*

As can be seen from the table, Downing's framework based on the *Standard's* framework covers all aspects of the 12-test development framework, although this was not a straightforward process. However, when looking at the Cambridge framework and comparing it to the other two in table 6.42, a number of points were not mentioned, or explicitly clear in the framework presentation. This demonstrates a gap in the Cambridge framework that may lead test developers to miss important sources of validity evidence when being used, and lack of reporting on this evidence. One important issue that was not clearly outlined in most of these frameworks was external factors that could be related to either culture, religion, natural disaster or lack of resources. Although external-related factors were not addressed within the Cambridge Validity

Framework for this study, the researchers did adjust processes for administering assessments by avoiding scheduling of examinations in the Holy month of Ramadan when all Muslims fast. Table 6.43 demonstrates further examples where the researcher was able to navigate and prepare for such events before the examinations.in order to minimise any effect due to these factors.

Table 6.43: External issues faced that may affect test development validity

Issues that were Faced During Test Development	Possible Threats that were Faced due to External Factors	How the Problem was Addressed
Issues that may affect test security	Cultural-factor: Identifying candidates for authentication because of the way they dress (long modest clothing or covering of the face)	Females were taken to a private area for authentication, and searching for materials before examination.
Sudden interruptions in the process and in gathering validity evidence	Natural disaster factor: A communicable disease during winter season prevented EM residents from gathering for discussions, and prevented exam committees from meeting for the item writing process in a timely manner as they were either mobilized to other hospitals or the hospital was partially closed for a precautionary quarantine measure.	Residents were re-scheduled and were given enough time to meet for discussions that suited them all and after the quarantine was ended and the environment was safe for all. Most of the item writing process started well before these events, and, therefore, re-scheduled meetings did not affect the flow of item development. The tests were completed, reviewed and delivered months before the actual testing schedule.
CIV that may affect the performance of examinee	Religion-factor: Fasting during the months of Ramadan, or various Islamic days throughout the year where Muslims fast occasionally may affect examinee performance.	Testing dates are usually scheduled a year in advance where the exam committee with the researcher avoided the month of Ramadan and other days to conduct an examination. An appropriate date was selected for these exams.
CIV that may affect the	Lack of resource: experts in the field of	The researcher's supervisory team was formed of members who were medical educators working in

response process, internal structure and scoring of the examination and test results	assessment in the region/country, as well as difficulty in finding psychometricians to aid in analysing the results	the field of assessment, as well as a psychometrician. In addition, the researcher travelled to neighbouring countries to meet with psychometricians and assessment expert to discuss issues related to test validity and psychometry. Furthermore, multiple workshops and conferences on validity, statistics, and psychometry were attended by the researcher to be able to understand and analyse the data.
Validation of the testing process and the research process	Lack of resource: The practice of conducting examinations and research studies frequently without following systematic guidelines to ensure the validity of the process and research. This is usually due to lack of knowledge or human resources to test developers.	The researcher aware of the importance of validity in testing utilized experts in the field and based the research on applying multiple validity frameworks to ensure that the intended interpretation of test results was valid as demonstrated. In addition, a further step was taken to legitimate the research process.

6.5 Conclusion

This concludes the results section for validity and validity frameworks. The results of the multiple sources of evidence gathered, the strength and weaknesses faced throughout the testing process and ways to overcome them, provide evidence of validity of the MM-TXT testing process. Quantitative results from pilot and test item analysis showed that multimedia items were more discriminating and took longer to answer than text items. Difficulty levels of both formats were similar and no significant differences were found. Results from the questionnaire outlined six main themes relating to multimedia items and issues related to CBT. Focus group thematic analysis yielded seven themes related to multimedia quality and characteristics, CBT, clarity and difficulty of the question, time and miscellaneous factors. The multiple results that were presented in this chapter are further discussed in the following chapter (Discussion), where results from both quantitative, as well as qualitative methods were pooled together (mixed-methods) to demonstrate the effect of multimedia on MCQ items in testing. It also covers the applicability of the Cambridge Validity Framework and views related to its process.

Chapter 7: Discussion

7.1 Introduction

The recent development in assessment procedures in Saudi Arabia led to exploring new methods to assess physicians in high-stakes examinations. The previous chapters explored the characteristics and effectiveness of multimedia by reviewing published literature, which was mostly stemmed in the learning environment. This research further compared multimedia to text-matched items using the Cambridge Validity Framework that was investigated. The quantitative set up of this research aimed at finding the effect of multimedia items on MCQs scores and item statistics in comparison with text questions as a means of measuring cognitive skills. While the qualitative aspect acts as an explanatory method for understanding the quantitative data.

To the best of the author's knowledge, this is the first study in Saudi Arabia that investigated the use of multimedia in high-stakes written examinations based on international guidelines and validity frameworks.

7.2 General overview of CBT and multimedia examination

CBT was viewed by most residents as a nice and comfortable experienced favoured over PNP exams, as demonstrated by Figure 6.16. The most appreciated expressed benefit of the multimedia (MM) examination was that the items reflected real-life cases as dealt with in practice (section 6.3.1.4.1). The MCQs used in the research stressed diagnostic information-gathering through reading the history, examination, and lab results in addition to the multimedia. Most residents preferred the MM format to their usual paper-based examinations, as demonstrated in Section 6.2 (Figures 6.12-6.15). The results coincide with previous studies (55, 80). This could be explained by the congruency effect, which explains that when information is encoded in a picture-based learning style, it is probably better retrieved in a picture-based testing setting (121).

Computer features personalize the experience for the test taker and offers a means of test strategies for examinees as expressed by the resident (use of highlights and mark item) to have a more comfortable and better way of answering the items and play an important role in the design of the examination, which is also documented in the literature (152, 165 p 339).

7.2.1 Item difficulty

Mean test scores for text groups (M= 75.3) were slightly higher than those of the multimedia group (M = 73.4) but were not significant. Although results were not significant, looking at the items individually, it does imply that most text items seem to be more direct and probably easier than multimedia questions. Both formats had a moderate difficulty level (Diff = 0.75,0.74), which is considered optimal for item discrimination. The correlation for difficulty between formats was high (r=0.81), which indicated that the items that were difficult in the text format were also difficult in the multimedia format.

Although the same range of difficulty levels in both formats was equal, there were items that behaved differently between formats. For example, when a multimedia item was more difficult (less in mean test score) items were probably not clear and had quality issues as indicated by the residents (Section 6.3.1.1). This might have had an effect on students' ability to identify the relevant findings and, hence, their performance (55). Another explanation why some multimedia items might have been perceived more difficult was because it was more authentic and did not give out any hints, as did the text version (e.g., description of an ultrasound). This goes with what residents said that text items gave away the answer and required no further thinking from them. In addition, in one of the studies comparing both formats, description of audio-clips in testing was

easier and gave out hints more than the audio format (20). More so, multimedia items were probably testing higher levels of cognition and so were perceived harder and were documented in the literature (45).

On the other hand, when an item was less direct in the text format (e.g., description of physical examination of a cut finger) the multimedia item made it easier to understand. The study by Shen and Li (2010) explained that content experts for the multimedia items believed that when narrating a situation or movement in the text format, the information became less direct and that candidates could gain richer information from the multimedia version that demonstrated a certain level of “feel” and “look” to them (5). This was demonstrated by the example of tendon injury in the hand that was easier for residents to view in the video more than the text (Figure 5.3, Chapter 5). However, multimedia was more difficult when it replaced textbook terminology, therefore, explaining why sometimes multimedia items were perceived easier or difficult according to the students’ perception of it. Similar results were found in the Shen et al. study (5).

In addition, some residents felt that although text items had the description, it required them to imagine the clinical situation, X-ray, or imagine the ECG report, which required more cognitive skills and according to cognitive load theory an additional mental process; therefore, making the text item more difficult (Section 3.3.7.3.2). Constructing a mental model from text alone requires some effort and may be subjected to misinterpretations and affect one’s comprehension because the text would need to be interpreted, unlike pictures where the mental model is directly formed (121). This is in line with Vorstenbosch et al’s. (2013) study where students needed to interpret text information translating it into images (49). It was also noted from the discussion (6.3.1.1) and literature, why some multimedia items were easier than text and was due to either the picture not being related to the content and the item could be solved without it or because

it was so clear that it acted as a cue and gave away the answer (121). However, conflicting views from other residents expressed that, when the multimedia had no relation to the item, it made them confused, required extra effort from them to make a link, and wasted their time. This was also reflected by students in another study (39). They also felt that some multimedia lacked labels, had they been present it would have made the item more helpful (39). This could be explained by the cognitive load theory, when redundant information is present, it has a negative effect and increases the cognitive load in the working memory by processing unnecessary information (124, 135, 139). In addition, Cook (2006) explained in his article about visual representation, and that the use of graphics may turn out to be functionally useless to the learner, even though it was designed to be cognitively useful (114). This is because the learner did not perceive the information in the intended manner that the graph was originally designed for. Vague or unclear multimedia for whatever reason may lead to construct irrelevant variance. This has been highlighted by other studies (99, 152) and affects test result interpretation and validity. As commented by the residents (Chapter 6), some items were more difficult because the MM did not correlate with the image. The reason is that the image did not fit with the verbal material and mental integration; therefore, would be difficult. Examinees' attention would be split between two modes of information as also explained in the modality principle (114) and based on the CLT, the split-attention effect occurs causing overload on the working memory (39, 126, 139).

In addition, there seems to be discrepancies between what consultants perceive would be a difficult, medium, and easy item for residents and what residents actually performed on the exam. This research demonstrated that while experts had classified items to be difficult or not based on the content, residents had another perspective of what was difficult or not based on the presentation and clarity of the items given. For instance, as

demonstrated in Chapter 6, Table 6.21 while consultants had labelled 14 items to be of the most difficult level for residents only one was classified as difficult by IA of residents' responses. The rest were either of medium difficulty or easy items. While by item analysis calculations, there were actually 14 difficult items but they were not the same ones that were labelled by the consultant as being difficult. From the item analysis, four of the items that were labelled as easy by the consultant and nine that were labelled as medium were actually difficult for the residents. This could be due to the items in some instances being very clear and unambiguous making those having the knowledge answering it hence, rendering the item as easy and not difficult for the resident. In addition, it could be that what consultants assumed to be difficult for the residents based on content or domain was not true for residents, as was because of how they think like experts. Experts have a rich knowledge base, need shorter viewing times and know where to best look for abnormalities through their rich knowledge base foundation (150) while novice learners tend to have pieces of information that are weakly connected, also known as fragmented knowledge (i.e., constraint to surface feature) (114).

Overall, labelling items by cognition levels (K1 or K2) seemed to align and reflect the residents' performance more with the items than the consultant's judgment. This was demonstrated by the presence of no differences between difficulty index calculated by item analysis and item cognition level set by the reviewers (Chapter 6 Table 6.23).

Results also showed that most but not all questions were difficult no matter what the format was, as shown in Table 6.16. This would reflect that the content was perceived with the same difficulty but the format may be testing a different element of the construct being measured, hence fluctuating the difficulty levels between text and multimedia formats.

Furthermore, the complexity of the item (having more results, labs and multimedia) doesn't necessarily indicate that the item would be more difficult. On the contrary, results have shown that what consultants have labelled difficult items and what reviewers have labelled as items having higher cognitive skills, residents perceived as easy. This does not indicate that there were no higher cognitive functions taking place; it just indicates that put all together, the items might have been so clear as presented in a real-life setting that whoever had the information no matter how complex or simple the item was would get it correct. And the opposite was true. Some items that were labelled as easy and as recall questions turned out to be difficult to the residents. This was mostly due to the multimedia being irrelevant to the question, not relating to the content or wasn't a classical or exaggerated presentation as expected by the residents.

Another explanation for the failure of visual resources can be attributed to the residents' learning styles as mentioned by them (Chapter 6), as not all residents were adequately able to process the images equally. Depending on the item format and personality of the residents, items seem to be interpreted differently when taken in the multimedia or text format. Preferred learning style plays a role in how the items are perceived by the candidates. Some perceive visual media as challenging and stimulate cognitive skills while others see it as recalling an image from practice or a book and is considered a spot diagnosis with no thinking required. This was also noted by students in a study where using graphics, audio, video and animations helped them recall information. In addition, in the speciality of nursing examination with an AV format, nurses felt that the visuals helped them recall information and hence, they spent more time understanding the concept and less time trying to memorize the details (80) while others perceive the same for text items. Some residents perceived written descriptions of a condition as

challenging requiring imagination and thinking while others who preferred to read description found it to be as recalling information.

Visual skills and clinical experience also play an important role in how residents interpreted the data from multimedia and depend on knowledge and experience as pointed out by the residents. Residents' level of knowledge and expertise in the EM content area may have affected their performance with multimedia (38); the less prior knowledge one had, the more likely they were to be subjected to cognitive overload (114). Prior knowledge can determine how easy a learner can interpret and perceive visual representation in working memory (114). Therefore, visual skills varied between training years among residents and were reflected in their scores. This was apparent in test results where mean tests scores in both formats showed a statistically significant difference among residency levels (demonstrated in Table 6.13 Chapter 6). This was more apparent between R1s and R2s and R3s. This is supported by studies in radiological expertise where visual skills development depended on knowledge base and clinical experience (42). These results (seniors scoring higher than juniors) support the construct validity of the test items.

In addition, individual characteristics such as prior knowledge, cognitive skills, spatial ability and learning style preferences affect how residents interact with the item (38, 50, 114). In the case of ultrasound and echo items, residents found it more difficult to interpret the image, as well as the video than the text version. This was identified because of the short loop of video and not enough labelling or information given regarding the surrounding structures, as well as the image format was being viewed as a standard two-dimensional (2D) image instead of the usual three-dimensional (3D) view in clinical practice. In this situation, by the time the residents viewed the relevant pictures from the video clip the next segment re-looped, not giving them enough time for deeper

processing (139). This led to cognitive overload because the cognitive capacity was not enough to cope with the processing demands (139). This is all related to the anatomy of the structure, special orientation and spatial relations for example, of the uterus to other organs, sacs and spaces. This is important in order to be able to understand what is being viewed and is understandable as, during residency training, postgraduates interact with the patient and conduct radiological procedures that are of multiple views (54). A similar example was explained in the speciality of anatomy where viewing the carpal bones was in the anterior and posterior view in the atlas while in the multiple views in the dissection lab (147).

There were no differences in test scores by gender or region between residents. This indicated that both genders and all regions were receiving the appropriate level of training that was set by SCFHS training programs. There were differences between level of residents in mean test scores. This was expected and should be the case as the higher the residency training, the more knowledge the residents would have and the higher mean test scores were expected. This also provides strength to validity evidence.

7.2.2 Discrimination, point biserial and reliability

Multimedia questions were significantly more discriminating than their text-matched questions (p =.03), Table 6.16. Furthermore, discrimination was higher in the multimedia items particularly in the medium/difficult group; however, it was not significant. This was also seen in the cross-tabulation results in Table 6.29 where results showed that the majority of items in the text format were of the recall/understanding type (K2-C) and those of the multimedia items were of the higher cognitive levels (K2-A/B). This also reflected in the consultants' judgment regarding what residents would perceive as being difficult. The text format had almost equal distribution of easy and medium/difficult items

based on the consultants' labelling, while the multimedia format contained mainly of medium/difficult items as displayed in Table 6.30. In addition, residents perceived that the exam required them to do more than recall facts as reflected in the survey results (Figures 6.12-6.15). There were two studies that did not have similar results, regarding images having no uniform effect on item discrimination (54, 55). The reasons for this remain unexplained (54).

From the residents' feedback, the discrimination of the items seemed to vary according to the question and its content, in addition to the various factors that they have outlined. This goes with the weak correlation that was found between the discriminating index for both groups (r=0.23) in Figure 6.9. Sometimes, items were more discriminating in the text format and other times it was more discriminating in the multimedia format. This could be due to the nature of the visual data that may have an effect on mean test scores and, hence, item difficulty and discrimination. Hunt noted that students' performance in different areas varied when categorising items according to visual types. Items containing chest radiographs that required interpretation had higher discriminating levels while items that displayed fracture images, were more difficult and had no improvement in their discrimination indexes (55).

The point biserial was also higher for the multimedia group than the text with a borderline significance. This is expected with a higher discrimination for the multimedia group. The high discrimination and point biserial for the multimedia items indicate a better correlation of items in the multimedia format to the whole examination. The higher the correlation, the more reliable the exam is. This was true when calculating the reliability coefficient and G-coefficient for the multimedia and text items. Overall, the multimedia format had a higher reliability than the text format in the pilot items and similar reliability to the text items in the main study. This indicated that reliability was not lost by changing testing

formats. Reliability estimates suggest that reliability in well-selected multimedia items and with enough numbers improve reliability. This agrees with another study that found no change in reliability among test formats in a computerized test using AV items in the speciality of nursing (5). The G-coefficient calculated highlighted that residents' scores did not only vary because of individual differences in their knowledge or measured skills but also due to the items and their interaction with the items as expressed through their comments in the focus groups. A higher reliability can be achieved using these higher fidelity questions in CBT, as they closely resemble the actual clinical stimuli than the text questions (70). In addition, the larger the number of items that are included in an examination, the more reliable the test scores would be (32 p.129).

7.2.3 Duration

On average, MM items took about five seconds longer to complete than text items as shown in Table 6.16, Chapter 6. Although this difference may seem small on an individual item basis, it was statistically significant ($p < 0.01$). This is important, as it may have a significant impact on the total time needed to answer items particularly if most or all items in an examination were augmented with MM. This was also implied by Bersky (1994) (72). This also explains the time restriction felt by residents to complete the items as reflected in the survey (Figure 6.12). The complexity and length of the items allowed residents to complete the items but not to have enough time to review the items.

Although multimedia items took longer to answer, there was a high correlation with the text format ($r=.74$) in Figure 6.11, which indicated that questions that took longer to answer in the multimedia format also took longer to answer in the text format. The longer duration spent on the multimedia item could be explained by it being more complex and requiring more time. In addition, some items were perceived easy but due to the quality

of the multimedia, it took longer to look at or it was required to repeat the video. This was also reflected in a study that pointed out that lengthy and frequent inspection of images does not mean an improvement in learning results (123). Residents explained that the longer duration on items was due to having to spend more time on an image to understand and search for the clinical signs, repeating the videos more than once because of the short loop, orienting themselves to the anatomy of the structures seen and due to the difficulty level of items. This was reflected in the correlation between item difficulty and duration where 18% of the durations could be explained by the difficulty level.

In some instances, multimedia items were answered quicker than text items. Three explanations were given by the residents as: a) the multimedia was very clear making the item easy and so less time was spent on reading or trying to remember how something exactly looked like and, therefore, there was more time to focus on the features of the multimedia and what it meant in order to be able to apply it (80); b) the duration spent on the items did not always give an accurate reflection on their thinking process. Some items were too difficult and were answered quickly for any answer, instead of wasting more time on something they knew they weren't going to answer; and c) some descriptions of procedures or signs such as ECG were felt more complicated or less direct than its multimedia pair and took longer to understand and imagine.

In addition, the duration spent on an item and the number of items an examinee can answer per minutes depend on different factors, such as the type of question and media, complexity of the required thought process to answer the item, and the examinees' test habits (32 p.129). In addition, in CBT, the actual time spent on multimedia items is also explained by the type of item and the nature of the task (165 p.343). It is apparent that

multimedia items should have more time allowed to answer based on the results and this would strengthen the validity argument and justify the additional testing time (20).

7.2.4 Characteristic of multimedia items

Multimedia items as reflected in the questionnaire (Figure 6.13-6.15) and from the emerging themes of the focus group (Section 6.3.1.4) highlights the characteristic of multimedia being realistic and reflecting clinical practice more, containing more information about the patient's situation compared to text and aiding in confirming resident's diagnosis by making their decisions more certain (Figure 6.13). In a study conducted by Crisp and Sweiry (2006), students felt that the presence of an illustration reassured them. However, this depended on how students' perceived images as being useful and helpful (120).

Residents expressed that multimedia items either acted as an addition to the item, provided information that was usually not described in the text format and provided the whole description of the clinical presentation. As described by Peeck (1993), multimedia can be seen as an adjunct to the main text, as well as a significant source of information on its own. However, this is only when it is adequately extracted for proper interpretation (123).

Multimedia's effects on examinees' performance go beyond its surface characteristics. As it involves applying cognitive abilities (clinical reasoning), physical spatial and perceptual skills (e.g., pattern recognition, visual search, and visual information processing), as well as features that interact with the examinee. This was also reflected by Ravesloot et al. (2012) and Parshall, Davey and Pashley (2000) (42, 95). However, this interaction is not fully understood until now.

7.2.5 Item format

In general, MM items could be either difficult or easy, and the differences in difficulty were likely related to the item format and to the amount of the information that the MM content added to or took away from the test items relative to their text-based counterpart. This was also demonstrated by Linjun Shen et al. 2010 (5), where multimedia resources were most useful when they served a specific purpose (e.g., supplying additional information or clarifying a concept or procedure), and when included they must be clear, contain minimal irrelevance, complement the corresponding text and do not cause ambiguity (120).

Residents felt that some features of multimedia were missed when given in the image format when compared to text. Such features were related to the brightness of images, having more than one cut of radiological image (e.g., CT) scan. As cross-sectional images require a set of consecutive images to scroll through in order to make a diagnosis and greatly differs from looking at a single image. This is backed up by studies that suggest that these types of images are extraordinary with respect to cognitive process and visual information (31, 50, 150).

In addition, residents expressed that multimedia items provided specific information that was not usually translated in the text format (e.g., height of the ECG waves, its relation to other waves, background colour of skin bedding, the absence of certain signs, the direction of a rash and the actual severity of the clinical condition). Multimedia items also provided information that were hindering to test performance by providing images or videos that contained other signs that were not relevant to the case (e.g., presence of runny nose and type of skin acne in a foreign body case, the presence of examiner's

hand in an image that confused the residents, ear covered by hair in an ear infection question).

It is a fallacy to consider all multimedia (either images or videos) as being equal. It seems highly possible that the effect of multimedia depends on its relevance to the context of the item, and depends on the type of image or video used (40). These were all expressed by the residents throughout their discussions (Section 6.3.1).

7.2.6 Multimedia, Higher-Cognitive Levels and Validity

It is known from the literature (Chapter 3), that the use of multimedia may test a different element of a construct (cognitive powers) than the text version of the item, and this can only be demonstrated if the two forms perform differently (5, 77). That means that the context of the item has substantially changed and resulted in each format measuring a different construct (5). This was demonstrated in the results of this research in item parameters, particularly item discrimination and duration.

Results from item analysis showed that multimedia items were more discriminating, took longer to answer and, to some extent, were more difficult and had a higher point biserial (Section 6.1.2.2). Residents also expressed multiple times that the presence of these items stimulated them to think and use higher cognitive levels (Section 6.3.1.4.4).

When multimedia makes an item more difficult than the text version, then this could either be explained by the presence of cues in the text version or the requirement of an extra skill by the examinee to interpret the cues found in the multimedia (54). When an extra skill is required from a well written clear multimedia item this meant that it has an impact on item validity, and adds value to the role of multimedia in measuring the construct being tested (54, 229). Such extra skills could be occurring in the working memory as explained by the cognitive load theory (CLT). We can also assume that the

skills needed to extract information to reach the correct answer from multimedia are different from those used to extract it from the text format. (54). This suggests that images have an effect on test validity (54).

In written examination, validity is limited by its low fidelity for testing physical examination findings because tasks presented in written questions are dissimilar to real-life experiences. Improving the authenticity of the question by making the stimulus more realistic and similar to the actual setting (e.g., by showing rather than describing the abnormality) should improve the validity of score interpretation (70). Using words and describing physical findings in scenarios bypasses the important skill of identifying and interpreting the abnormality in a given case. Replacing what is described with actually showing the physical abnormality using multimedia makes the question more resembling to the actual application of knowledge or skill in the clinical setting (70). This strengthens the validity of score interpretation. Much research is needed in this area, as until present few data support this proclamation.

The authenticity that is added by the use of multimedia is not enough to be included in licensing examination if no additional measurement value is added, or if it distracts the examinee from the intended construct (5). However, the results of previous studies demonstrated that the use of multimedia items are promising in that they are capable of measuring elements of the construct differently than those of the text items (5).

The results of this study are consistent with results from a study on the use of innovative questions to improve assessment of nursing practice; that demonstrated that these items were more difficult and better discriminating than their counterpart in the text format, students felt these items were more representative of their actual work performance and that the video items required more cognitive skills and were perceived as more authentic

than their matched MCQ items (45). The results are also consistent with the literature that well-written questions, enhanced with multimedia that are aimed to test problem-solving and interpretation skills require a higher level of cognitive performance or an additional skill more than questions with a written description of the scenario (45, 55).

7.3 Validity framework

This study tried to clarify and conceptualise how the Cambridge framework fits within the Saudi context of assessment and tried to provide a framework for medical educators not only in developed countries but also in developing ones that may face different challenges. Analysis of the framework tried to determine the evidence, methods used, as well as the challenges faced and gaps for assessment that needed further elaboration. The framework, as well as other frameworks, did not contain answers on how to deal with situations when interruptions occurred during test development that would affect the validity process; or when certain steps of the framework could not be completed fully. The framework was found to be difficult to understand and the language not simple for someone outside the field of assessment. In addition, the researcher needed to explore other frameworks, such as the APA framework described by Downing (164), Kane described by Cook (157) and the framework for effective test development; which was more comprehensible (168). Although the framework is theoretical, practical applications in the clinical setting referring back and forth to the various steps of the framework is highly needed. Furthermore, from the literature review, the studies with their various methods and limitations, none had described the in-depth evaluation of the validity framework used, and few got direct feedback from their examinees through the use of surveys only. From the literature, there was no systematic way described for educators or clinicians to follow the framework practically and reporting it with examples,

except for three articles. Two by the same author Andreatta (2009, 2011). One article was on validity evidence in simulation article which tackled a specific simulation procedure and mapped it against validity evidence for further explanation (37), and the other one was about validity in a competency-based assessment in obstetrics and gynaecology (156). The third article was by Cook, Brydges, Ginsburg and Hatalas (2015) and was on the approach to validity argument. As the authors presented a practical guide to Kane's framework (157). As it should be noted that the language used in Kane's report on the framework was not always easy to follow and comprehend.

Most of the studies also did not clearly discuss the reporting of validity evidence, which is important in judging the evidence (160), the literature still lacks behind in reporting the quality for evaluating assessment tools. And although a lot of guidelines are present, the task of reporting felt challenging with no practical examples to follow.

7.4 Limitations

Like all research, this research has several limitations that should be considered when interpreting the results.

7.4.1 Research method constraint

Despite the strength of mixed-method research, it does have its limitations. Limit to generalizability in mixed-methods are due to the sampling techniques. For the quantitative method, generalization cannot directly be applied from the population because the sample was randomly assigned but not randomly selected. However, this could be overcome by replicating the study with different individuals at different times and at different places. This is known as the "replication logic" (159 p.269). For qualitative method, because participants in the focus groups were not randomly selected

and the sample size may sometimes be considered small, researchers must be careful when generalizing results from focus groups (159 p.239).

7.4.2 Sample size constraint

The sample size of the items and residents may be viewed by some as not being large enough (164 residents combined and 80 items combined). The lacking number of questions investigated was as a result of resource constraints, which may limit the external validity of this research. Unfortunately, developing and analysing a large number of multimedia questions can be difficult, as such research relies on a number of stakeholders and their cooperation. In addition, it would be extremely difficult to include a large number of these questions during a single examination. Regarding residents' number, the whole population was included. Although the number of multimedia–text item pairs might not be sufficient to make a confirmatory conclusion, the number of questions examined in this research does provide insight to the use of multimedia items as a method of testing higher cognitive skills.

The study did not include the audio type of media as explained in Section 5.3.3.6.3.3, as it wasn't fitting for the speciality of EM in the format of MCQs. Most types of audio (conversations, heart sounds, breath sounds etc.) are either examined during EM residents' clinical examination, or are more commonly used in other sub-specialities. For example, heart sounds are tested more in the speciality of Cardiology, communication and consultation conversation in the speciality of psychiatry, breath sounds and wheezes in the speciality of pulmonology and so forth. In addition, the studies in the literature review mostly referred to videos and images and to some extent, audio.

Therefore, results may not be generalizable to specialities that use more of this type of

format. In addition, results in this study were not correlated with another form of assessment method testing the same construct to ensure validity.

This multimedia format would need to be tested on other specialities to determine the "fit" for purpose and generalizability of results and any additional requirements or improvements that need to be made. Although extensive time and review went into developing the items, the testing format could still be improved and become more user-friendly as suggested by the residents (being able to change image resolution, providing more than one cut image). These changes may further improve the exam results and reliability (109). However, the researcher believes that the overall exam process and results were not greatly affected by any one of these limitations.

7.4.3 External variables that affect results

In this study, even though residents were randomly assigned to different forms of questions, different instructors taught the different residents according to their regions, so there was no control over the effectiveness of the instructional process. However, reliability and item analysis demonstrated that there were no differences among residents by gender or region. Although residents were informed of the study beforehand, some didn't receive the follow-up notice and felt uncomfortable. This and the effect of other factors such as stress, fatigue, and loss of concentration could not be determined. This examination was based on SCFHS policies and did not include a break; however, after the residents' comments a break was implemented in SCFHS long examinations. Some testing procedures were not applicable, optimal or according to the framework (e.g., the use of well-known standard-setting techniques, the use of IRT theory in analysis). This depends on the criteria set by the policymakers. Classical test theory (CTT) is mostly used in the Middle East probably because of it being more

understandable and easier to implement, as well as the lack of experts in IRT methods in the region. In addition, IRT was not applicable in this research as explained in (Section 5.3.3.7.1) because of the sample size, which requires a large number. Therefore, DIF was also not possible. In addition, DIF could not be done, as all the participants were Saudi and of the same culture and religious background. However, CTT and G-theory and statistical analysis were applied to residents' demographics, which looked at mean group differences by level, gender and region, as well as differences in items in relation to content, difficulty, cognition, level, gender, region and type, with no differences found. This ensured fairness of the items. Moreover, focus group discussions did not reveal unfair items from the residents' perspectives. Finally, fairness and item bias prevention was also covered through standardised testing condition, scoring procedures, and statistical characteristics that provide evidence for validity and reflect test takers' knowledge of the assessed construct (155, 168)

7.5 Reflection

7.5.1 The researcher as an instrument in design, data collection and analysis

This section covers the role of the researcher as she carries out the research process. The researcher has a bachelor's degree in medicine and surgery as an undergraduate followed by a master's degree in medical education. Having started up the Department of Medical Education with a Senior Medical Educator with expertise in assessment, the researcher's focus and interest in working in SCFHS were mainly in the area of assessment. Working in the SCFHS and meeting various physicians from all specialities and different regions gave insight on what to expect. Having worked with the EM committees since they first joined also gave the researcher familiarity with the

EM setting and gave insight on what to expect as challenges and expectations for the residents. Being involved in conducting the workshops and reviewing items also gave confidence and insight on what to expect and review in the study in order to ensure the quality of the delivered items, as well as to look for factors that may act as a CIV.

Using the validity framework shortly after the development of the Assessment unit in the Commission gave the researcher time to take a critical look and reflection and compare the steps in the framework with the set-up of the department. Because the use of frameworks searches for validity threats, it also gave an opportunity to reflect on any negative evidence that may have risen that were against the research conclusion, and theoretical explanation. Self-reflection and overcoming biases were considered, for example, through listening to the recording and taking a second look at the researcher's justification, responses to some issues, and looking at it from another perspective.

The research design used was a mixed-method design drawing on different research methods. In the qualitative aspect of the research, the researcher was considered an instrument of data collection as the researcher was the one asking the questions, collecting, and analysing the data (159 p.36). Therefore, the researcher was aware of this point when trying to capture the participants' viewpoints to have a better understanding of what was felt by the residents. In qualitative research, this is referred to as empathetic understanding (159 p.36). During the focus group (FG), no observer was present because of resource constraints and unfamiliarity with the process and during the discussion, the researcher was aware of two things: 1) That most or probably none of the participants have ever been in a focus group before, and, therefore, did not know what to expect; and 2) because of the nature of the Saudi population it is not common amongst Saudis to express what they feel. The culture and the way most are brought up even in schools do not focus on reflecting and expressing one's feelings.

Therefore, the researcher was aware that this might be a challenge and that some participants would feel conscious or shy during this process. In addition, it was recognised that the role of the researcher as head of assessment would make residents intrigued to open up subjects regarding their examinations during the focus group. Therefore, the following points were used in order to ease the tension and facilitate the group discussion:

1) The FG started with a brief introduction of the researcher and participants

2) Explanation of what was the purpose of the meeting, and what was expected, and

3) Unrelated to the project, participants were allowed time to express what they desired in regards to their examinations.

Participants felt more comfortable speaking and interacting after they had expressed their concerns and had their questions answered and were able to better focus on the FG discussion. The researcher being a female moderator in a more masculine field didn't feel that there was an impact during the focus group. The impact of gender on data generation is not that straightforward, as other factors usually come into consideration such as demonstrating sensitivity and expressing genuine interest by the moderator (209 p.50). During the discussion, the researcher took notes and during the analysis phase, all of which played a final role in interpreting the results. During the transcription phase, to minimise errors, transcription of the whole focus group conversations was done by the researcher. One of the ways to reduce error is for the transcriber to have a background of the context and subject matter that is being transcribed, and to understand the accent, pace and rhythm of the participants. And, in this case, have a further skill in medical terminology, as well as understanding and speaking both the languages of Arabic and English (220 p.18). Regarding quantitative analysis, it was objective, generated by the computer on well-defined and approved

standards and criteria in assessment. Results were discussed and reviewed with the supervisory team, statisticians and medical educators in the field of assessment and helped to make the judgments on the results sound.

An important aspect to be mindful of were the sources of knowledge that helped shape the thinking and process of the research and research study. The sources of knowledge are concerned with the study of the theory and justification of knowledge also known as epistemology (159 p.12). The study used was a combination of rationalism as the inductive reasoning method with empiricism and the deductive (confirmatory) method and pragmatism as a way of combining methods to find solutions, solve problems, and answer the questions. The researcher also had an influence on how the study, design, data collection and interpretation were dealt with. The sources of beliefs of the researcher needed to be identified as it shaped how one learns from the world around him/her. Sources of belief can stem out of family and friends, tradition, culture and religion, books, as well as thinking and experiencing the world. As a Saudi, Arab, Muslim female researcher there is no doubt that the sources of knowledge would be different from someone else with a different religion or backward from somewhere else. However, it is also the combination of these resources that also helped shape and give trustworthiness to the research. As an Arab generally, and a Saudi specifically, the culture and family around the researcher helped in building a character that is community-driven, embracing honesty and the search for truth. As a Muslim, it is at the heart of Islam to practise honesty, kindness, altruism, beneficence, and doing the best in everything in one's daily life. In addition, it is also in the heart of our practice to have faith and believe in things that are not seen or felt. Table 5.1 in Chapter 5 described an overview of the epistemology and ontology taken by the researcher. What one believes in what validity is and what it constitutes is what determines his/her viewpoint on what is

considered as validity evidence. It is apparent that this belief is connected to the culture that one lives in and his/her epistemological belief on knowledge and how it is found and constructed. This, in turn, is tied to some views to religious beliefs and how one acquires knowledge. Validity can be viewed as an integrative process that requires high measurement precision as a piece of evidence. At the same time, from an orthogonal viewpoint, validity can be viewed as a separate property of psychometric functioning. For example, it can be viewed that high reliability is not necessary nor sufficient for validity; as one may have an instrument that measures the construct but has a low measurement precision (158). In regards to the researcher, the basic well-known concept of validity was taken. However, it is viewed from a point of degree, and the absence of some evidence whether because of lack of resources, sample size or unawareness doesn't entail the results to be totally invalid. Meaning, if all proper steps are taken in a rigours manner; however, the results turn out to be unrelated, does not mean that the results are not valid.

The ethical stances that the residents have voiced were valid and considered a threat to validity. However, this was beyond the researcher's role or position to change. Like anywhere else, each organisation sets its own policies and procedures that fit their culture, economic and political environment. The only thing one can do when test factors collide with what the researcher wishes are to minimise its effect and try to implement change.

7.5.2 Framework

In general, the structure of the framework seems to be simple. However, its simplicity proved to be challenging as it did not provide detailed information on what to do in every step. Even with the reviewed edition of the framework that came out after the research was conducted. The review of the framework was mostly linguistic (230). Moreover, in addition, factors that seemed to hinder this research process and affect its flow were first viewed as threats to validity. However, when revising the framework, it was noticed that out-of-hand factors were not mentioned in the framework and were not explained on how to deal with. This led to the view that these were areas that were missed and that there was a gap that needed to be reviewed. Therefore, the idea of including these factors into the framework as an additional section that could influence the flow of test development and affect countries more than others. This is where the notion of international validity was explored. The Cambridge framework tends to be applicable for international examinations, which may assume that the construct for a test is similar across culture. However, it was difficult to assume that the measured construct, as well as behaviours, are identical across cultural groups, how much was overlapped and what were the social component of the instruments in the framework. Culture is associated with ethnicity, religion, regional aspects and specific institutions (197). The ethnic composition of Saudi citizens is mostly Arab, religion is Islam, and the regional aspects are related to how the community is set up. The institution depends on the qualification and experience of related policymakers familiar with assessment, as well as the availability of resources in these areas. The Cambridge framework, as well as other frameworks, did not cover aspects of test administration that have an effect on results.

An example of cultural issues that might affect the validity of test results and their

interpretations by introducing threats to validity or CIV, is during test security and administration that was faced was the cultural dressing of the female (wearing a fully covered dress (abaya) and face cover(niqab)). This may make authentication difficult in a mixed environment where no private secure room was available at the time of exam setting. In addition, identifying any hidden materials in pockets under the dressings, as well as searching females in front of their male colleagues would be a major issue of disrespect and invasion of privacy. Exam centre outline was not all designed for this purpose in the first place and so lack of place is a physical issue (189). Another example of issues faced were sudden interruptions to gathering validity evidence appropriately by natural or manmade disasters that were beyond the control of the researcher and test developers. For example, during the preparation of the examination, a spread of certain diseases during winter season affected the examination process. Precautionary measures were taken, such as partial hospital closure and preventing residents from different hospitals and staff to gather. This affected meeting times and item writing flow process. In addition, another incidence was a sudden lack of trained experts that were shifted to another region in the country where patient care was needed (e.g., most physicians of various specialities travel to Makkah during the Hajj pilgrimage as a national and moral duty, and fast during the Holy month of Ramadan). This has an effect on the test process and timing of planned events (e.g., preparing materials, publishing blueprints, developing items), etc. Although most of the time exams are not scheduled during these months, testing cannot be stopped for these two whole months. In the month of Ramadan where fasting is required, this may affect examinee's performance. The breaking of fast is when the sun sets and so, examination is difficult to be conducted at that time, particularly when it is a holy month of worship and prayers. Therefore, the threats of

validity in these situations are unclear in how they affect the results and inferences made from these results. Although these issues might be seasonal or occasional and might affect certain cultures, regions, or countries more than others, they are still important issues that need to be cleared, understood and addressed during testing and in the researcher's opinion should be added to validity frameworks, to be aware of such issues and prepare beforehand when applicable, particularly when a great deal of time and effort is put in conducting a structured examination. Such preparation could be, designating a private area for security and searching female candidates before entering an exam, blocking certain dates or months for testing when a religious event is expected, preparing materials beforehand, and starting exam preparation earlier to avoid any delay in testing.

Issues that were faced during the whole process that was not covered in any framework were issues related to education research, politics, and decision making that were inextricably intertwined. The following points were left unanswered by the researcher:

- How to deal with imbedded cultural and religious events and issues that are justified when they affect certain aspects of test conduction (e.g., Hajj, Ramadan)
- Although validity is a degree and not all of it can be applied to one setting, but are there minimum acquired standards to have as a basis for a valid test, where one works their way from the ground up increasing its strength?
- In addition, when validity evidence is gathered appropriately but is interrupted by factors that are beyond the control of test developers (e.g., spread of disease, lack of experts, religious events) and have an effect on the timing of planned events (e.g., preparing materials, publishing blueprints, developing items, etc),

what is the state of validity in these situations, particularly when a great deal of time and effort would have been put in?

- When a change in organisational leadership and a turn-over of policymakers and decision-makers govern the organisation and with each policymaker, new rules are adopted and others are dropped irrespective of it affecting the validity framework. This is mostly because of dealing with stakeholders that are not familiar with the concept of test validity and the detailed process of quality assurance needed. As resistance to change plays an important role in introducing threats to validity during testing.
- Lack of experts in specific fields (assessment experts, psychometricians with backgrounds in G-theory and IRT theory), as the main practice in the Kingdom is CTT.

Countries greatly differ in the degree of policies, regulations and legal control they exercise in testing. In addition, depending on the resources available (e.g., assessment experts and psychometricians) organisations set their own standards and mechanisms for controlling them in testing (e.g., analysis done by a third party or non-psychometricians). International validity may have a place if the appropriate framework was selected that also included challenges that may be faced in different cultures and populations. This is to minimise CIV and to ensure that inferences reached from test results hold up in other populations that are linguistically, culturally, and socioeconomically different (199) This is achieved in exported assessment, and, therefore, perhaps exported validity frameworks could be developed. Exported assessments as stated by Oliver, Lawless and Young (199) is the assessment that ‘are developed in one country and are used in countries with a population that differs from the one which the assessment was developed’. An example of this is the Graduate

Record Examinations (GRE) to make a decision for higher education admission. In addition, with the availability of International Test Commission (2001), the International Guidelines for Test Use and International Journal of Testing (82) list guidelines on test use that are needed at an international level. However, they do highlight that contextual conditions must be considered at the local level when being implemented and that these conditions may affect how the guidelines may be realised and managed in practice. The conditions include social, political, institutional, linguistic and cultural differences between assessment settings. It also includes legal documentation, local laws applied to countries, international standard that addresses testing issues, as well as national guidelines (189). The same view might be implemented to validity framework.

Furthermore, the move towards CBT and online testing evidenced by the work being carried out in the USA and UK raises a whole host of issues related to testing process security and administration that may have a place in the framework. The idea of exported assessment, International ICT program, International Test Commission [ITC] (2005, 2013), as well as the (*Standards for Educational and Psychological Testing)* (American Educational Research Association [AERA], American Psychological Association [APA], and the National Council on Measurement in Education [NCME], 2014) all play a role in developing new and improved frameworks.

7.6 Conclusion

MCQ with multimedia are more discriminating and take longer to answer than the classical text MCQs. Their difficulty depends on the type, characteristic, and fit of the multimedia with the accompanying text. The use of multimedia items rather than text increases the construct and face validity of the test questions. Multimedia items improve assessment validity by engaging students in an authentic clinical task that resemble their actual practice and that elicit their knowledge and higher cognitive skills. The development of a MM exam is an iterative process, requires multiple viewing, feedback and revision by professionals. Because it can be difficult to conceptualize and describe the concept while using multimedia, seeing the material is essential and feedback from examinees is vital.

Developing a computer-based exam using multimedia is a lot more than simply converting a paper-based exam to a computer-based one. This is due to the fact that multimedia materials are more dynamic and are very different than what is used in the conventional method.

Before simply adopting a new or existing assessment method, instrument or framework, one must first question and think about the appropriateness and effect/ influence of these measurements form a philosophical and conceptual point of view as it is being conceptualised and operationalised in a different culture or setting and whatever similarities are shared in these instruments may not be generalized or guaranteed across populations (197). Using an instrument, method or measure and applying it from one setting and culture to another requires attention to the cultural relevance. Therefore, there is no one best framework that is applicable for all, even if the framework has been described by a regulatory body. Consumers of test validity theories and frameworks

should focus more on the content instead of form and be familiar with the multiple vocabularies that validity carries. Producers should know the differences between validity theories and definitions and focus on arguments that limit possibility theories rather than on more definitions and frameworks. Different scholars, thinkers and groups will need to wear their own sunglasses to view validity. They will need to decide on their own filtering process on what entails as being important to them and in their own context.

7.7 Recommendation and Future research

The literature demonstrates that there is little valid and reliable research that can offer guidelines for test developers and organisations that wish to carry out the development of computer-based examinations using multimedia (39, 94, 231). All information gained from the literature studies and this research were compiled in a recommendation table (Table 7.1) that contains factors to consider when selecting multimedia materials that may affect multimedia and should be considered by all test developers. In addition, Table 7.2, describes characteristics and the nature of multimedia materials that were organised into three headings with factors related to the multimedia material itself and software interface, the accompanying text that should have a clear relationship with the MM and, finally, the individual's characteristics (123). These should be used to inform examination committees, as well as in the training item writers to be used as evidence to strengthen exam validity (120).

To strengthen the results of this research, further correlational studies and group performance should be carried out to correlate the results of MCQ examinations enhanced with multimedia with other formats of written examinations, as well as clinical examinations (e.g., OSCE). In addition, conducting a think-aloud protocol for a test with these item formats are necessary to gain deeper insight into the resident's cognitive

processes of thinking and interpreting the items with multimedia and improving the quality and development of selected materials. As most research is focused on the use of multimedia for instructional design and learning. Another area that needs further research is in the complexity and types of multimedia used (i.e., videos, CT, clinical image, X-ray, etc.), as well as their features (multiple cut images of CT images, resolution changes of radiological images) and their effects in the items and on examinees' cognitive process. All these would inform the development of guidelines for their use in examinations.

Finally, there is a place for research on what is the best way to use validity frameworks and report them practically using simple language that is understood to those who are novice in the test development process. In addition, the elaboration on unanticipated challenges (natural and man-made) that affect the validity process and the possible addition to existing frameworks as a means of exploring the addition of international validity (a framework that can be used by all).

Table 7.1: Factors to consider when selecting multimedia materials

Material Selection	Explanations
Type	It is important to properly select multimedia materials for problem-solving examinations (55) by using different types of multimedia (still, dynamic, both or none) for different purposes, fields and content domains (38, 124, 140).
Content	Multimedia cases need to be submitted by experts in the speciality, and in accordance with general guidelines of the blueprint (19). Selected video clips should allow the student to grasp the whole rather than the detail of the clip to avoid focusing on details that require multiple viewing (122).
Relevancy	Multimedia materials that are selected as test items should be relevant and relay the correct information needed (49)
Quality	It is important to consider the quality of the multimedia used when selecting the item so it does not be distracting (5).
Collaboration	Content experts should work closely with those providing multimedia material in order to ensure high-quality materials (56).
Required Skills and Knowledge	When selecting videos for item constructions, subject matter experts and educationalist should work together to have a combined knowledge of video technology, source credibility, subject content, current theory in multimedia learning and best practices of using video (124)
Material Orientation	It is recommended when selecting images or videos to understand and make clear the interaction between the multimedia and spatial ability as this interaction is expected during assessment (54, 147).
Expectations	Students' expectation regarding what element in a question is considered relevant should be considered when writing the item, to avoid the MM receiving more attention than the information provided from the text (120). Additional time should be factored in, to allow students to carefully view and review the multimedia according to the selected multimedia (122, 140, 144).

Table 7. 2: Nature of multimedia materials

1. Variables Related to the multimedia material itself
• **Measurement:** length, size, duration, position, volume (5, 56, 124). Interaction (MM and special ability) (124, 126).
• **Formatting issues:** Number of materials per item, line length, number of lines, layout on screen, screen resolution, screen size, font style and size, interline spacing, white space, sound, clarity, quality, pixilation, scrolling (5, 39, 56, 124).
• **Angle and Appearance:** Side (right, left), view, frames, rotation, positioning, different cut levels, angles, single or stack view, correct exposure (5, 56, 58, 80).
• **Colour Versatility:** Colour (presence or absence), resolution, shading (highlight and shadow detail), light exposure, fine-tuning of contrast, tonal range, brightness control (39, 56, 58, 123).
• **Demographic (related to MM material):** age, gender, level, region, ethnicity (39).
• **Dimensionality**: two-dimensional materials (illustrated) or three-dimensional material (real) (38, 147).
• **Fidelity:** the degree to which the item and media represents the context of a real-world situation (authenticity) (95, 99)
• **Interactivity:** the degree to which the item responds to examinees' input. For example, select answer format, simple feedback after response, or examinee interacts with the multimedia gaining more information with each interaction (38, 45, 97, 99, 118).
• **Level of abstraction:** realistic representation (visible concept), or abstract representation (non-visible concept) (38).
• **Complexity:** level of difficulty, number of tasks the examinee must take and consider (99)
• **Type of Media:** static (illustrations, photos, graphs, charts or maps) or dynamic, (audio, video, animated or interactive illustrations) (95, 99, 139)
• **Item Format:** multimedia used in stimulus or response format (118, 232), selected-response or constructed response (95, 118).
• **Input devices:** Keyboard, mouse, headphone, touch screen, light pen, joysticks, speech recognition software (59, 95).
• **Response Action:** physical actions examinees need to take in order to respond to the item e.g., clicking, typing, navigating the screen, drag-and-drop, pull-down menu, touch screen, speak into the microphone (59, 95, 99, 118).
• **Mode of Presentation:** paper, software or PowerPoint (145).
• **Assessment Structure:** display individual item or by group, situated task or simulated task (39, 99).
• **Information:** presented as narrated, text or image (95), Duplication (in MM and text or only one format).
• **Exclusive rights:** Sources permission and copyright issues of multimedia (7), new or used multimedia (40).
• **Other factors:** Location, landmarks identification (58), patient information needed or removed (19), calculator requirement, measurement tools (e.g., ruler, ECG calibrated paper, lab results), navigation tools (pause, stop, loop), on-screen text, closed captioning, audio-only or video only items (7, 39, 94, 123, 124).
2. Variables Related to the accompanying Text
• Relationship with the text content and context (123, 143).
• Degree of abstraction and situation described in the text content and context (38, 143).
• Difficulty level of the accompanying text scenario (39, 123).
• Information presented as narrated, text or with an image (95), Duplication (in MM and text or only one format).

- Density of information (simple or complex) in the material, and organisation of the information in the material (123).

3. Variables Related to the Individual

- Prior Knowledge, memory and comprehension (39, 123)
- Age, gender, ethnicity and level (39, 123)
- Ability, reading skills, language and visual literacy (123)
- Motivational factor (39, 99, 120, 122)
- Stress and anxiety (39, 63, 80, 102).
- Rate of reading from on-screen text (39)
- Familiarity with the material and computer interface use (39, 80, 94, 100, 143).
- Learning style preference (120)

References

1. Round J, Conradi E, Poulton T. Improving assessment with virtual patients. Medical Teacher. 2009;31(8):759-63.

2. Tarrant M, Ware J. A framework for improving the quality of multiple-choice assessments. Nurse educator. 2012;37(3):98-104.

3. Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. Academic medicine. 1990;65(9):S59-60.

4. Froncek B, Hirschfeld G, Thielsch MT. Characteristics of effective exams—Development and validation of an instrument for evaluating written exams. Studies in Educational Evaluation. 2014;43:79-87.

5. Shen L, Li F, Wattleworth R, Filipetto F. The promise and challenge of including multimedia items in medical licensure examinations: Some insights from an empirical trial. Academic Medicine. 2010;85(10):S56-S9.

6. Schuwirth LWT, Van Der Vleuten CP, De Kock CA, Peperkamp AG, Donkers HH. Computerized case-based testing: modern method to assess clinical decision making. Medical Teacher. 1996;18(4):294-9.

7. Bennett RE, Goodman M, Hessinger J, Kahn H, Ligget J, Marshall G, Zack J. Using multimedia in large-scale computer-based testing programs. Computers in Human Behavior. 1999;15(3–4):283-94.

8. Humphris GM, Kaney S. The objective structured video exam for assessment of communication skills. Medical Education. 2000;34(11):939-45.

9. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. BMJ. 2010;341:c5064.

10. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. Medical Teacher. 2016;38(1):3-17.

11. Kim MK, Patel RA, Uchizono JA, Beck L. Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. American Journal of Pharmaceutical Education. 2012;76(6):114.

12. Glaser R. Education and thinking: The role of knowledge. American Psychologist. 1984;39(2):93.

13. McCoubrie P. Improving the fairness of multiple-choice questions: A literature review. Medical Teacher. 2004;26(8):709-12.
14. Morrison S, Free KW. Writing multiple-choice test items that promote and measure critical thinking. Journal of Nursing Education. 2001;40(1):17-24.
15. Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? Academic Medicine. 2002;77 (10):S103-4.
16. Miller GE. The assessment of clinical skills/competence/performance. Academic Medicine. 1990;65(9):S63-7.
17. Harlen W, James M. Assessment and learning: Differences and relationships between formative and summative assessment. Assessment in Education: Principles, Policy & Practice. 1997;4(3):365-79.
18. Van Der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. Advances in Health Sciences Education. 1996;1(1):41-67.
19. Shea JA, Norcini JJ, Baranowski RA, Langdon LO, Popp RL. A comparison of video and print formats in the assessment of skill in Interpreting cardiovascular motion studies. Evaluation & the health professions. 1992;15(3):325-40.
20. Holtzman KZ, Swanson DB, Ouyang W, Hussie K, Allbee K. Use of multimedia on the step 1 and step 2 clinical knowledge components of USMLE: a controlled trial of the impact on item characteristics. Academic Medicine. 2009;84(10):S90-3.
21. Azer SA. Assessment in a problem-based learning course: Twelve tips for constructing multiple choice questions that test students' cognitive skills. Biochemistry and Molecular Biology Education. 2003;31(6):428-34.
22. Bailey PH, Mossey S, Moroso S, Cloutier JD, Love A. Implications of multiple-choice testing in nursing education. Nurse Education Today. 2012;32(6):e40-e4.
23. Su WM. Writing context-dependent item sets that reflect critical thinking learning outcomes. Nurse Educator. 2007;32(1):11-5.
24. Ilgen JS, Humbert AJ, Kuhn G, Hansen ML, Norman GR, Eva KW, Charlin B, Sherbino J. Assessing diagnostic reasoning: A consensus statement summarizing theory practice and future needs. Academic Emergency Medicine.2012;19(12):1454-61.

25. Freiwald T, Salimi M, Khaljani E, Harendza S. Pattern recognition as a concept for multiple-choice questions in a national licensing exam. BMC Medical Education. 2014;14(1):232.
26. Downing SM, Baranowski RA, Grosso LJ, Norcini JJ. Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical specialty certification. Applied Measurement in Education. 1995;8(2):187-207.
27. Haladyna T. Writing Test Items to Evaluate Higher Order Thinking. Boston,MA: Allyn & Bacon;1997.
28. Heist BS, Gonzalo JD, Durning S, Torre D, Elnicki DM. Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: A mixed methods study. Journal of Graduate Medical Education. 2014;6(4):709-14.
29. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. The Lancet. 2001;357(9260):945-9.
30. Van Bruggen L, Manrique-van Woudenbergh M, Spierenburg E, Vos J. Preferred question types for computer-based assessment of clinical reasoning: A literature study. Perspectives on Medical Education. 2012;1(4):162-71.
31. Van der Gijp A, Van der Schaaf MF, Van der Schaaf IC, Huige JCBM, Ravesloot CJ, Van Schaik JP, ten Cate TJ. Interpretation of radiological images: Towards a framework of knowledge and skills. Advances in Health Sciences Education. 2014;19(4):565-80.
32. Achievement test Planning In: Ebel RL, Frisbie DA. Essentials of Educational Measurement. 5th ed. New Delhi: Prentice hall of India; 2009: p.126-132
33. Content and Cognitive Processes In: Haladyna TM. Developing and validating multiple-choice test items. 3rd ed. Routledge; 2015: p.20-27
34. Atkins MJ, O'Halloran C. AMEE Medical Education Guide No. 6. Evaluating multimedia applications for medical education. Medical Teacher. 1995;17(2):149-60.
35. Shaw S, Crisp V, Johnson N. A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assessment in Education: Principles, Policy & Practice. 2012;19(2):159-76.
36. Kane M, Crooks T, Cohen A. Validating measures of performance. Educational Measurement: Issues and Practice. 1999;18(2):5-17.
37. Andreatta PB, Gruppen LD. Conceptualising and classifying validity evidence for simulation. Medical Education. 2009;43(11):1028-35.

38. Ruiz JG, Cook DA, Levinson AJ. Computer animations in medical education: A critical literature review. Medical Education. 2009;43(9):838-46.
39. Hao Y. Does multimedia help students answer test items? Computers in Human Behaviour. 2010,26(5):1149-1157.
40. Holland J, O'Sullivan R, Arnett R. Is a picture worth a thousand words: An analysis of the difficulty and discrimination parameters of illustrated vs. text-alone vignettes in histology multiple choice questions. BMC Medical Education. 2015;15(1):184.
41. Frank J, Snell L, Sherbino J. CanMEDS 2015 Physician competency framework. Ottawa: Royal College of Physician and Surgeon of Canada; 2015.
42. Ravesloot C, van der Schaaf M, Haaring C, Kruitwagen C, Beek E, Ten Cate O, van Schaik J. Construct validation of progress testing to measure knowledge and visual skills in radiology. Medical Teacher. 2012;34(12):1047-55.
43. Tayal VS, Centers S, Snead G, Norton HJ. 273: Evaluation of emergency medicine resident introductory ultrasound rotation by a multimedia testing competency tool. Annals of Emergency Medicine. 2007;50(3):S86.
44. Peterson MW, Gordon J, Elliott S, Kreiter C. Computer-based testing: Initial report of extensive use in a medical school curriculum. Teaching and Learning in Medicine. 2004;16(1):51-9.
45. Wendt A, Harmes JC. Developing and evaluating innovative items for the NCLEX: Part 2, item characteristics and cognitive processing.Nurse Educator.2009;34(3):109-13.
46. Bond WF, Spillane L, CORD Core Competencies Simulation Group. The use of simulation for emergency medicine resident assessment. Academic Emergency Medicine. 2002;9(11):1295-9.
47. Okuda Y, Bryson EO, DeMaria S, Jr., Jacobson L, Quinones J, Shen B, Levine AI. The utility of simulation in medical education: What is the evidence? Mount Sinai Journal of Medicine. 2009;76(4):330-43.
48. Ilgen JS, Sherbino J, Cook DA. Technology-enhanced simulation in emergency medicine: a systematic review and meta-analysis. Academic Emergency Medicine. 2013;20(2):117-27.
49. Notebaert AJ. The effect of images on item statistics in multiple choice anatomy examinations. Anatomical Sciences Education. 2017;10(1):68-78.
50. Vorstenbosch MA, Bouter ST, Van den Hurk MM, Kooloos JG, Bolhuis SM, Laan RF. Exploring the validity of assessment in anatomy: Do images influence cognitive

processes used in answering extended matching questions? Anatomical Sciences Education. 2014;7(2):107-16.
51. Schuwirth LW, Van der Vleuten C, Donkers HH. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
52. Biggs J. Aligning teaching for constructing learning. The Higher Education Academy. 2003:1(4).
53. Newhouse CP. Using digital technologies to improve the authenticity of performance assessment for high-stakes purposes. Technology, Pedagogy and Education. 2015;24(1):17-33.
54. Vorstenbosch MA, Klaassen TP, Kooloos JG, Bolhuis SM, Laan RF. Do images influence assessment in anatomy? Exploring the effect of images on item difficulty and item discrimination. Anatomical Sciences Education. 2013;6(1):29-41.
55. Hunt DR. Illustrated multiple choice examinations. Medical Education. 1978;12(6):417-20.
56. Buzzard AJ, Bandaranayake R, Harvey C. How to produce visual material for multiple choice examinations. Medical Teacher. 1987;9(4):451-6.
57. Saudi Commission for Health Specialities [Internet]. About SCFHS. 2012 [updated 11/13/2012; cited 2018 Sep 1]. Available from: https://www.scfhs.org.sa/en/about/pages/organization.aspx.
58. Akhtar S, Theodoro D, Gaspari R, Tayal V, Sierzenski P, LaMantia J, Stahmer S, Raio C. Resident training in emergency ultrasound: Consensus recommendations from the 2008 Council of Emergency Medicine Residency Directors Conference. Academic Emergency Medicine. 2009;16 (2):S32-6.
59. Sireci SG, Zenisky AL. Innovative item formats in computer-based testing: In pursuit of improved construct representation. Handbook of Test Development. 2006, 18:329-47.
60. Case SM, Swanson DB., & National Board of Medical Examiners. Constructing written test questions for the basic and clinical sciences. (Revised). Philadelphia, PA: National Board of Medical Examiners; 2002.
61. Wood T, Cole G, Lee C. Developing multiple choice questions for the RCPSC certification examinations. Ottawa (ON): Office of Education, Royal College of Physicians and Surgeons of Canada, 2010.
62. Paniagua MA, Swygert KA.Constructing written test questions for the basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners. 2016.

63. Akdemir O, Oguz A. Computer-based testing: An alternative for the assessment of Turkish undergraduate students. Computers & Education. 2008;51(3):1198-204.
64. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. [Internet] PLOS Medicine. Public Library of Science; 2009. Available from: http://doi.org/10.1371/journal.pmed.1000097.
65. Santos CM, Pimenta CA, Nobre MR. The PICO strategy for the research question construction and evidence search. Rev Lat Am Enfermagem. 2007;15(3):508-11.
66. Cantillon P, Irish B, Sales D. Using computers for assessment in medicine. Bmj. 2004;329(7466):606-9.
67. Boulet JR, Swanson DB. Psychometric challenges of using simulations for high-stakes assessment. Simulations in Critical Care Education and Beyond. Des Plains, IL: Society of Critical Care Medicice 2004; Dec:119-30.
68. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. Bmj. 2005;331(7524):1064-5.
69. Fakhoury H, Hajeer A. Guide to Literature Search for Medical Students. In: Al Alwan I, Magzoub ME, Elzubeir M, editors. International Handbook of Medical Education: A Guide for Students. London: Sage; 2012;p.95-99.
70. Lieberman SA, Frye AW, Litwins SD, Rasmusson KA, Boulet JR. Introduction of patient video clips into computer-based testing: Effects on item statistics and reliability estimates. Academic Medicine: 2003;78(10):S48-51.
71. Buzzard AJ, Bandaranayake RC. Comparison of the performance of visual and verbal multiple-choice questions. Australian and New Zealand Journal of Surgery. 1991;61(8):614-8.
72. Bersky AK. Effect of audiovisual enhancement on problem-solving and decision-making activities during a computerized clinical simulation test (CSTTM) of nursing competence. Evaluation & The Health Professions. 1994;17(4):446-64.
73. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. Journal of Health Services Research & Policy. 2005;10 (1_suppl):6-20.
74. Bugbee JrAC. The equivalence of paper-and-pencil and computer-based testing. Journal of Research on Computing in Education. 1996;28(3):282-99.

75. Mills CN, Potenza MT, Fremer JJ, Ward WC, editors. Computer-based testing: Building the foundation for future assessments. London: Routledge, Taylor & Fracis Group; 2018.
76. Craig N. Mills, Maria TP, John J. Fremer, William C. Ward. Computer-Based Testing. New York: Routledge. 1st ed. New York, 2002.
77. Cheng I, Basu A. Improving multimedia innovative item types for computer based testing. In eighth IEEE International Symposium on Multimedia,2006. p. 557-66.
78. Burr SA, Chatterjee A, Gibson S, Coombes L, Wilkinson S. Key points to facilitate the adoption of computer-based assessments. Journal of Medical Education and Curricular Development. 2016;3:77-83.
79. Ellaway R, Masters K. AMEE Guide 32: E-Learning in medical education Part 1: Learning, teaching and assessment. Medical Teacher. 2008;30(5):455-73.
80. Liu M, Papathanasiou E, Hao YW. Exploring the use of multimedia examination formats in undergraduate teaching: Results from the fielding testing. Computers in Human Behavior. 2001;17(3):225-48.
81. Tunc Y, Armstead M. Computer based testing: The ball state experience. In Proceedings of the 29th annual ACM SIGUCCS conference on User services; Portland, Oregon, USA. 2001, p. 201-3.
82. International Test Commission. International guidelines for test use. International Journal of Testing. 2001;1(2):93-114.
83. International Test Commission. International guidelines on computer-based and internet-delivered testing. International Journal of Testing. 2006;6(2):143-71.
84. Zhang X, Cai F, Liu F, Bao X, Liu Y. A new and cheap medical examination system with artificial intelligence. Medinfo. 1995;8:1702.
85. Dillon GF, Clauser BE. Computer-delivered patient simulations in the United States medical licensing examination (USMLE). Simulation in Healthcare. 2009;4(1):30-4.
86. Lim EC, Ong BK, Wilder-Smith EP, Seet RC. Computer-based versus pen-and-paper testing: Students' perception. Annals Academy of Medicine Singapore. 2006;35(9):599.
87. Wei H. Computer-based testing (CBT) and the USMLE. Medical Computing Today. 1999;9(4):267-9.

88. Dillon GF, Clyman SG, Clauser BE, Margolis MJ. The introduction of computer-based case simulations into the United States medical licensing examination. Academic Medicine. 2002;77(10):S94-6.
89. Irish B, Sales D. Does computer-based testing (CBA) have a future in the assessment of general practitioners in the United Kingdom? Education for Primary Care. 2006;17(1):1-9.
90. Guagnano MT, Merlitti D, Manigrasso MR, Pace-Palitti V, Sensi S. New medical licensing examination using computer-based case simulations and standardized patients. Academic Medicine. 2002;77(1):87-90.
91. Lievens F. The ITC Guidelines on computer-based and internet-delivered testing: where do we go from here? International Journal of Testing. 2006;6(2):189-94.
92. Lin H, Dwyer F. The fingertip effects of computer-based assessment in education. TechTrends. Linking Research and Practice to Improve Learning. A Publication of the Association for Educational Communications & Technology. 2006;50(6):27-31.
93. Hols-Elders W, Bloemendaal P, Bos N, Quaak M, Sijstermans R, De Jong P. Twelve tips for computer-based assessment in medical education. Medical Teacher. 2008;30(7):673-8.
94. Schoech D. Using video clips as test questions: The development and use of a multimedia exam. Journal of Technology in Human Services. 2001;18 (3-4):117-31.
95. Parshall CG, Davey T, Pashley PJ. Innovative item types for computerized testing. In Computerized Adaptive Testing: Theory and Practice, Springer, Dordrecht: 2000. (p. 129-48).
96. Noyes JM, Garland KJ. Computer- vs. paper-based tasks: Are they equivalent? Ergonomics. 2008;51(9):1352-75.
97. Strain-Seymour E, Way WD, Dolan RP. Strategies and processes for developing innovative items in large-scale assessments: Pearson Education, Inc.;2009. p.21.
98. Wendt A, Harmes JC. Evaluating innovative items for the NCLEX, part I: Usability and pilot testing. Nurse Educator. 2009;34(2):56-9.
99. Bryant W. Developing a strategy for using technology-enhanced items in large-scale standardized test. Practical Assessment, Research & Evaluation. 2017;22(1):10.
100. Schoenfeldt LF. Guidelines for computer-based psychological tests and interpretations. Computers in Human Behavior. 1989;5(1):13-21.

101. Al-Saleem SM, Ullah H. Security considerations and recommendations in computer-based testing. The Scientific World Journal;2014:7.
102. Boevé AJ, Meijer RR, Albers CJ, Beetsma Y, Bosker RJ. Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. PloS one. 2015;10(12).
103. Russell M, Goldberg A, O'Connor K. Computer-based testing and validity: A look back into the future. Assessment in Education: Principles, Policy & Practice. 2003;10(3):279-93.
104. Chua YP. Effects of computer-based testing on test performance and testing motivation. Computers in Human Behavior. 2012;28(5):1580-6.
105. Karay Y, Schauber SK, Stosch C, Schüttpelz-Brauns K. Computer versus paper—Does it make any difference in test performance? Teaching and Learning in Medicine. 2015;27(1):57-62.
106. Boulet JR, Swanson DB. Psychometric challenges of using simulations for high-stakes assessment. Stimulations in Critical Care Education and Beyond. Des Plains, IL: Society of Critical Care Medicine. 2007,119-130.
107. Feinstein E, Gustavson LP, Levine HG. Measuring the instructional validity of clinical simulation problems. Evaluation & the Health Professions. 1983;6(1):61-76.
108. Hatala R, Kassen BO, Nishikawa J, Cole G, Issenberg SB. Incorporating simulation technology in a Canadian internal medicine specialty examination: A descriptive report. Academic Medicine: 2005;80(6):554-6.
109. Millos RT, Gordon DL, Issenberg SB, Reynolds PS, Lewis SL, McGaghie WC, Petrusa ER. Development of a reliable multimedia, computer-based measure of clinical skills in bedside neurology. Academic Medicine. 2003;78(10):S52-.5
110. Dictionaries O. Multimedia | Definition of Multimedia in English Oxford University Press; 2018 [cited 2018 May 2]. Available from: oxforddictionaries.com/definition/multimedia.
111. Schnotz W. Commentary:Towards an integrated view of learning from text and visual displays. Educational Psychology Review. 2002;14(1):101-20.
112. Dindar M, Yurdakul IK, Dönmez FI. Multimedia in test items: Animated questions vs. static graphics questions. Procedia - Social and Behavioral Sciences. 2013;106:1876-82.

113. Andrusyszyn MA. The effect of the lecture discussion teaching method with and without audio-visual augmentation on immediate and retention learning. Nurse Education Today. 1990;10(3):172-80.
114. Cook MP. Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. Science Education. 2006;90(6):1073-91.
115. Chinchor NA, Thomas JJ, Wong PC, Christel MG, Ribarsky W. Multimedia analysis + visual analytics = multimedia analytics. IEEE Computer Graphics and Applications. 2010;30(5):52-60.
116. Dyson MC. How physical text layout affects reading from screen. Behaviour & Information Technology. 2004;23(6):377-93.
117. Gulikers JT, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. Educational Technology Research and Development. 2004;52(3):67.
118. Parshall CG, Harmes JC. Improving the quality of innovative item types: Four tasks for design and development. Journal of Applied Testing Technology. 2014;10(1):1-20.
119. Jarodzka H, Janssen N, Kirschner PA, Erkens G. Avoiding split attention in computer-based testing: Is neglecting additional information facilitative? British Journal of Educational Technology. 2015;46(4):803-17.
120. Crisp V, Sweiry E. Can a picture ruin a thousand words? The effects of visual resources in exam questions. Educational Research. 2006;48(2):139-54.
121. Lindner MA, Eitel A, Barenthien J, Köller O. An integrative study on learning and testing with multimedia: Effects on students' performance and metacognition. Learning and Instruction. 2018.
122. Hertenstein MJ, Wayand JF. Video-based test questions: A novel means of evaluation. Journal of Instructional Psychology. 2008;35(2):188.
123. Peeck J. Increasing picture effects in learning from illustrated text. Learning and Instruction. 1993;3(3):227-38.
124. Dong C, Goh PS. Twelve tips for the effective use of videos in medical education. Medical Teacher. 2015;37(2):140-5.
125. Isham D. Developing a computerized interactive visualization assessment. Journal of Computer-Aided Environmental Design and Education. 1997;3(1).

126. Berends IE, Van Lieshout EC. The effect of illustrations in arithmetic problem-solving:Effects of increased cognitive load. Learning and Instruction. 2009;19(4):345-53.
127. Malone S, Brunken R. Assessment of driving expertise using multiple choice questions including static vs. animated presentation of driving scenarios. Accident; Analysis and Prevention. 2013;51:112-9.
128. Ebel R [Measuring educational achievement.] Essentials of educational measurement,2nd edition. Englewood Cliffs, N.J.: Prentice-Hall; 1972.
129. Ferland JJ, Dorval J, Levasseur L. Measuring higher cognitive levels by multiple choice questions: A myth? Medical Education. 1987;21(2):109-13.
130. Downing SM, Norcini JJ, Jr., Kangilaski R, Popp RL, Cheitlin MD. Still-frame simulations of cardiac motion studies: Validity evidence from the cardiovascular disease certification examination. Academic Medicine. 1996;71(1):S43-5.
131. Chenkin J, Lee S, Huynh T, Bandiera G. Procedures can be learned on the Web: A randomized study of ultrasound-guided vascular access training. Academic Emergency Medicine. 2008;15(10):949-54.
132. Norman MK. Twelve tips for reducing production time and increasing long-term usability of instructional video. Medical Teacher. 2017;39(8):808-12.
133. Taslibeyaz E, Aydemir M, Karaman S. An analysis of research trends in articles on video usage in medical education. Education and Information Technologies. 2017;22(3):873-81.
134. Hao Y. Does multimedia help students answer test items? Computers in Human Behavior. 2010;26(5):1149-57.
135. Sweller J, Van Merrienboer JJ, Paas FG. Cognitive architecture and instructional design. Educational Psychology Review. 1998;10(3):251-96.
136. Gupta M, Ingle GK, Malhotra R, Malhotra C, Lal P. Use of digital images in the undergraduate Community Medicine examination. South-East Asian Journal of Medical Education. 2011;5(1):45.
137. Brainerd CJ, Desrochers A, Howe ML. Stages-of-learning analysis of picture-word effects in associative memory. Journal of Experimental Psychology:Human Learning and Memory. 1981;7(1):1-14.
138. Ellaway R. Reflecting on multimedia design principles in medical education. Medical Education. 2011;45(8):766-7.

139. Mayer RE, Moreno R. Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist. 2003 1;38(1):43-52.
140. Kühl T, Scheiter K, Gerjets P, Gemballa S. Can differences in learning strategies explain the benefits of learning from static and dynamic visualizations? Computers & Education. 2011;56(1):176-87.
141. Carney RN, Levin JR. Pictorial Illustrations still improve students' learning from text. Educational Psychology Review. 2002;14(1):5-26.
142. Dindar M, Kabakçi Yurdakul I, Dönmez FI. Measuring cognitive load in test items: Static graphics versus animated graphics. Journal of Computer Assisted Learning. 2015;31(2):148-61.
143. Wu HK, Kuo CY, Jen TH, Hsu YS. What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. Computers & Education. 2015;85:35-48.
144. Issa N, Schuller M, Santacaterina S, Shapiro M, Wang E, Mayer RE, DaRosa DA. Applying multimedia design principles enhances learning in medical education. Medical Education. 2011;45(8):818-26.
145. Jamet E, Gavota M, Quaireau C. Attention guiding in multimedia learning. Learning and Instruction. 2008;18(2):135-45.
146. Lindner MA, Eitel A, Strobel B, Köller O. Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. Learning and Instruction. 2017;47:91-102.
147. Garg AX, Norman G, Sperotable L. How medical students learn spatial anatomy. The Lancet. 2001;357(9253):363-4.
148. Kozhevnikov M, Hegarty M, Mayer RE. Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers. Cognition and Instruction. 2002;20(1):47-77.
149. Chandler P, Sweller J. Cognitive load theory and the format of instruction. cognition and instruction. 1991;8(4):293-332.
150. Van der Gijp A, Ravesloot CJ., Jarodzka H., Van der Schaaf MF, Van der Schaaf IC, Van Schaik JP, Ten Cate TJ. How visual search relates to visual diagnostic performance: A narrative systematic review of eye-tracking research in radiology. Advances in Health Sciences Education. 2017;22(3):765-87.

151. Chang ACS, Read J. Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. System. 2013;41(3):575-86.
152. De Klerk S, Veldkamp BP, Eggen TJ. A framework for designing and developing multimedia-based performance assessment in vocational education. Educational Technology Research and Development. 2018;66(1):147-71.
153. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955;52(4):281-302.
154. Messick S. Validity. Research Report Series. Educational Testing Service; Princeton: New Jersey. 1987(2):i-208.
155. AERA, APA, NCME Standards for Educational and Psychological Testing. Educational Measurement: Issues and Practice.2014, 33(4):4-12.
156. Andreatta PB, Marzano DA, Curran DS. Validity: what does it mean for competency-based assessment in obstetrics and gynecology? American Journal of Obstetrics and Gynecology. 2011;204(5):384.e1-.e6.
157. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments:A practical guide to Kane's framework. Medical Education.2015;49(6):560-75.
158. Newton PE, Shaw SD. Disagreement over the best way to use the word 'validity' and options for reaching consensus. Assessment in Education: Principles, Policy & Practice. 2016;23(2):178-97.
159. Johnson RB, Christensen L. Educational Research : Quantitative, Qualitative, and Mixed Approaches. SAGE Publications, Incorporated; 2016.
160. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. Academic Medicine: 2013;88(6):872-83.
161. Slomp DH, Corrigan JA, Sugimoto T. A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. Research in the Teaching of English. 2014;48(3):276.
162. Koretz D. Making the term 'validity' useful. Assessment in Education: Principles, Policy & Practice. 2016;23(2):290-2.
163. Pangaro L, Ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Medical Teacher. 2013;35(6):e1197-e210.

164. Downing SM. Validity: on the meaningful interpretation of assessment data. Medical Education. 2003;37(9):830-7.
165. Impara JJC, Foster D, Downing SM, Haladyna TM. Handbook of Test Development. Hoboken: Taylor & Francis, 2006.
166. Kane MT. Validating the interpretations and uses of test scores. Journal of Educational Measurement. 2013;50(1):1-73.
167. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. 1995;50(9):741-9.
168. Lane S, Raymond MR, Haladyna TM, editors. Handbook of test development. New York: Routledge; 2016.
169. Kane M, Burns M. The argument-based approach to validation. School Psychology Review. 2013;42(4):448-57.
170. Rosenthal M. Qualitative research methods: Why, when, and how to conduct interviews and focus groups in pharmacy research. Currents in Pharmacy Teaching and Learning. 2016;8(4):509-16.
171. Schram AB. A mixed methods content analysis of the research literature in science education. International Journal of Science Education. 2014;36(15):2619-38.
172. Stalmeijer RE, McNaughton N, Van Mook WN. Using focus groups in medical education research: AMEE Guide No. 91. Medical Teacher. 2014;36(11):923-39.
173. Onwuegbuzie AJ, Leech NL. On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. International Journal of Social Research Methodology. 2005;8(5):375-87.
174. Onwuegbuzie AJ, Collins KM. A Typology of mixed methods sampling designs in social science research. The Qualitative Report. 2007;12(2):281-316.
175. Tuckett AG. Applying thematic analysis theory to practice: A researcher's experience. Contemporary Nurse. 2005;19(1-2):75-87.
176. Breen RL. A practical guide to focus-group Research. Journal of Geography in Higher Education. 2006;30(3):463-75.
177. Kitzinger J. Qualitative research. Introducing focus groups. BMJ: British medical journal. 1995;311(7000):299-302.
178. Cohen L, Manion L, Morrison K. Research methods in education. London and New York: Routledge, 2007.

179. Colbran S, Gilding A, Colbran S. Animation and multiple-choice questions as a formative feedback tool for legal education. The Law Teacher. 2016:1-25.
180. Van Teijlingen ER, Rennie AM, Hundley V, Graham W. The importance of conducting and reporting pilot studies: The example of the Scottish Births Survey. Journal of Advanced Nursing. 2001;34(3):289-95.
181. Paul Jones RWS, Diane Talley. Handbook of Test Development. Hoboken: Taylor & Francis; 2006.
182. Naeem N, Van der Vleuten C, Alfaris E. Faculty development on item writing substantially improves item quality. Advances in Health Sciences Education. 2012;17(3):369-76.
183. Ware J, Kattan T, Siddiqui I, Mohammed AM. The perfect MCQ exam. Journal of Health Specialties. 2014;2(3):94-9.
184. Haladyna TM. Developing and validating multiple-choice test items (3rd ed.): Lawrence Erlbaum Associates Publishers, Mahwah, NJ; 2004.
185. Begum T. A guideline on developing effective multiple choice questions and construction of single best answer format. Journal of Bangladesh College of Physicians & Surgeons. 2012;30(3):159-66.
186. Siddiqui I, Ware J. Test blueprinting for multiple choice questions exams. Journal of Health Specialties. 2014;2(3):123-5.
187. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Education Today. 2006;26(8):662-71.
188. Vander Schee BA. Test item order, level of difficulty, and student performance in marketing education. Journal of Education for Business. 2013;88(1):36-42.
189. The International Test Commission (ITC). The guidelines on the security of tests, examinations,and other assessments. International Journal of Testing.2016;16(3):181-204.
190. Schuwirth LW, Van der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher. 2011;33(10):783-97.
191. Engelhardt PV. An introduction to classical test theory as applied to conceptual multiple-choice test. Getting Started in Physics Education Research. 2009;2(1):1-40.
192. McGahee TW, Ball J. How to read and really use an item analysis. Nurse Educator. 2009;34(4):166-71.

193. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Medical Teacher. 2012;34(11):960-92.
194. Solano-Flores G, Li M. Generalizability theory and the fair and valid assessment of linguistic minorities. Educational Research and Evaluation. 2013;19(2-3):245-63.
195. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Medical Education. 2010;44(1):109-17.
196. Clifton SL, Schriner CL. Assessing the quality of multiple-choice test items. Nurse Educator. 2010;35(1):12-6.
197. Lee J, Jung DY. Measurement issues across different cultures. Journal of Korean Academy of Nursing. 2006;36(8):1295-1300.
198. Bosher S. Barriers to creating a more culturally diverse nursing profession: Linguistic bias in multiple-choice nursing exams. Nursing Education Perspectives. 2003;24(1):25-34.
199. Oliveri ME, Lawless R, Young JW. A validity framework for the use and development of exported assessments. Educational Testing Service.Princeton,NJ:2015.
200. Hicks NA. Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. Nurse Educator. 2011;36(6):266-70.
201. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Medical Education. 2008;42(2):198-206.
202. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. Advances in Health Sciences Education: Theory and Practice. 2002;7(3):235-41.
203. Nedeau-Cayo R, Laughlin D, Rus L, Hall J. Assessment of item-writing flaws in multiple-choice questions. Journal for Nurses in Professional Development 2013;29(2):52-7.
204. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP(UK) examinations. BMC Medical Education. 2010;10(1):40.
205. Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability theory: A practical guide to study design, implementation, and interpretation. Journal of School Psychology. 2014;52(1):13-35.

206. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. Medical Education. 2002;36(10):972-8.
207. Huang J. Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. Assessing Writing. 2012;17(3):123-39.
208. Parker A, Tritter J. Focus group method and methodology: Current practice and recent debate. International Journal of Research & Method in Education. 2006;29(1):23-37.
209. Flick U, Kvale S, Angrosino M, Barbour R, Banks M, Gibbs G, Rapley T. The Sage qualitative research kit: Using visual data in qualitative research.5 5.London: SAGE; 2007.
210. Wilkinson S. Focus group methodology: A review. International Journal of Social Research Methodology. 1998;1(3):181-203.
211. Kitzinger J. The methodology of focus groups: the importance of interaction between research participants. Sociology of Health & Illness. 1994;16(1):103-21.
212. Powell R, Single H. Focus Groups. International Journal for Quality in Health Care.1996,8(5) 499-504.
213. Reed J, Payton VR. Focus groups: Issues of analysis and interpretation. Journal of Advanced Nursing. 1997;26(4):765-71.
214. Jayasekara RS. Focus groups in nursing research: Methodological perspectives. Nursing Outlook. 2012;60(6):411-6.
215. Jamieson L, Williams LM. Focus group methodology: Explanatory notes for the novice nurse researcher. Contemporary Nurse. 2003;14(3):271-80.
216. Greenwood M, Kendrick T, Davies H, Gill FJ. Hearing voices: Comparing two methods for analysis of focus group data. Applied Nursing Research. 2017;35:90-3.
217. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3(2):77-101.
218. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nursing & Health Sciences. 2013;15(3):398-405.
219. Hydén LC, Bülow PH. Who's talking: Drawing conclusions from focus groups—some methodological considerations. International Journal of Social Research Methodology. 2003;6(4):305-21.
220. Gibbs GR. Thematic coding and categorizing In: Analyzing Qualitative Data. 2007;703:38-56.

221. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care. 2007;19(6):349-57.
222. Basit T. Manual or electronic? The role of coding in qualitative data analysis. Educational Research. 2003;45(2):143-54.
223. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Education Today. 2004;24(2):105-12.
224. Braun V, Clarke V. Thematic analysis. The Journal of Positive Psychology. 2012, 12:3, 297-298.
225. Yaghan MA. "Arabizi": A contemporary style of arabic slang. Design Issues. 2008;24(2):39-52.
226. Onwuegbuzie AJ, Johnson RB. The validity issue in mixed research. Research in the Schools. 2006;13(1):48-63.
227. Pallant J. SPSS survival manual: A step by step guide to data analysis using IBM SPSS, 6th ed. Maidenhead: Open University Press;2016.
228. Kane MT. Validating interpretive arguments for licensure and certification examinations. Evaluation & the Health Professions. 1994;17(2):133-59.
229. Schuwirth LW, van der Vleuten CP. Different written assessment methods: What can be said about their strengths and weaknesses?Medical Education.2004;38(9):974-9.
230. Shaw S, Crisp V. Reflections on a framework for validation–five years on. Research Matters: A Cambridge Assessment Publication. 2015;19:31-7.
231. Solano-Flores G, Wang C, Shade C. International Semiotics: Item difficulty and the complexity of science item illustrations in the PISA-2009 international test comparison. International Journal of Testing. 2016;16(3):205-19.
232. Coderre SP, Harasym P, Mandin H, Fick G. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. BMC Medical Education. 2004;4(1):23.
233. Brady A-M. Assessment of learning with multiple-choice questions. Nurse Education in Practice. 2005;5(4):238-42.
234. Moss E. Multiple choice questions: Their value as an assessment tool. Current Opinion in Anaesthesiology. 2001;14(6):661-6.

235. Su WM, Osisek PJ, Montgomery C, Pellar S. Designing multiple-choice test items at higher cognitive levels. Nurse Educator. 2009;34(5):223-7.
236.Goyal N, Aldeen A, Leone K, Ilgen JS, Branzetti J, Kessler C. Assessing medical knowledge of emergency medicine residents. Academic Emergency Medicine. 2012;19(12):1360-5.

Appendices

Appendix 1 Variables related to the research question

Variables related to the research question

Variable Type/role (159) p 38-42	Key characteristics	Example	Effect
Independent variable (IV)	A variable that is assumed to cause changes to another variable	Type of question (text or multimedia)	The type of question (IV) affects test scores (DV)
Dependent variable (DV)	The variable that is changed because of another variable	Test performance parameters (score, difficulty, discrimination, duration)	The type of question (IV) affects test scores (DV)
Mediating (intervening) variable	A variable that comes in between other variables and changes (moderates) the relationship between other variables	Cognitive thinking process due to the type of multimedia used	The type of question leads to processing information in the working memory. The mediating variables may affect this process.
Moderator variable	A variable that outlines the course by which one variable affects the other	Level of residents Region of residents Gender	Perhaps the relationship between the type of question and IA parameters changes according to the different level of residents and their knowledge. Or changes according to their region they are trained or to their gender differences.
Extraneous variable	Is a variable that competes with the IV to explain an outcome	Preferred learning style (visual, auditory, kinetic, reading), or test-taking strategies (test-wiseness and item flaws)	Perhaps the observed relationship between question type and test scores is due to the different learning styles of the participants (visual, auditory, etc) or resident's test-taking strategies.
Confounding variable	An extraneous variable that is not controlled for and maybe the reason why a particular result is found. It varies with the independent variable and influences the dependent variable	Stress, fatigue, environment (noise, temperature)	Perhaps the effect of stress, fatigue during testing or the temperature in the exam room has an effect on residents' performance.

Appendix 2. Summary of the 11 Studies that used MM-TXT matched items

Summary of the 11 Studies that used MM-TXT matched items

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Hunt, 1978 (55)	Two parallel tests of key feature single best-answer items (written items matched with an identical illustrated one) were taken in a final examination by 142 final year medical students (groups were randomly selected), one-week later students were invited to take the opposite format. The illustration occurred in the MCQ stimulus format (stem).	Illustrated format calls for interpretation of visual material to test cognitive skills at levels above the level of recall. If MCQ is enhanced with visual data (for interpretation) then: Item difficulty will increase and discrimination will improve.	70 items were identical in all aspects to the first group except that the Stem contained an illustration of actual data instead of being written (radiological images). These were printed in a separate booklet. Time for illustrated format (90 min.) because of the additional paperwork involved.	70 items with Stems containing a written description of clinically relevant findings of the visual data (e.g., radiologist's report of an X-ray). Items were based on clinical problems. Time for written format (60 min.)	Including visual data produced changes in DIFF and Disc in most items. 43 items were harder with the image format, 18 were easier and 9 showed no differences between groups. **DIFF**: Significant Increase in difficulty Index (i.e., item easier) in the illustrated format **DISC**: No consistent changes on discrimination for the examination as a whole.		Students in both groups praised the illustrated format to be more clinically relevant and most favoured this format; if images were improved to be introduced in their final certifying examinations(55). Problems with reproduction and selection of visual data were revealed. Some items gave paradoxical results for e.g., where an illustrated item performed much better than the text item, students who didn't answer it were of the middle and high performers in the test. With further inspection, most students interpreted the image incorrectly and got the correct answer for the wrong reasons.

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Anthony et al., 1991 (71)	77 triplet question, in three MCQ papers, were given in 12 Part1 FRACS[2] examinations between (1983-1989). The questions contained a visual question (X), an equivalent verbal question (Y), and another verbal question testing different information but the same content (Z).	Study's question as stated in the article was: 'Are visual questions more difficult, and better able to discriminate "good" from "weak" candidates, than corresponding verbal questions testing the same information, and different information?' (71), With related sub-categorized hypothesis.	Question (X) contains a visual trigger material, and another verbal question testing different material than (X) but in the same content area (Z) were given.	An equivalent verbal question to question (X) testing the same material to (X) was constructed this is called Question (Y). Questions X and Z appeared together in the same test, while questions Y were given in a separate year.	There were no significant differences in mean difficulty levels or discrimination levels among the pairs of questions X, Y and Z.		Results could be attributed to the small sample size that may have masked the real differences. However, they do not undermine what is obvious, the importance of visual recognition in the surgery speciality and responding to visual stimuli as done in clinical practice. Students preferred visual questions and considered it to be more clinical and more acceptable. These views agree with Hunt's findings.
					DIFF	**DISC**	
					No differences between pairs	No differences between pairs	

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Judy et. al, 1992 (19)	Two forms very similar in content for each CV[3] motion studies were created in two formats (print and video) and delivered in 16 centres. 392 Students in eight centres were randomly assigned to take either Echo form A (4 centres) or B (4 centres), and the other centres took the other motion studies. An entire form is composed of both a video subtest and a print subtest (about half of the cases were videos and the other half were print).	To demonstrate if print and video formats measured interpretive abilities in the same way.	High-quality cardiovascular videos, reviewed and selected by experts according to the blueprint. Candidates were situated as a group around a single monitor. They had the opportunity to view the study and mark the answers twice.	For each case in the print format, a small number of frames taken from an analogous cardiovascular video case in the other form was presented when possible. They were reviewed and selected by experts according to the blueprint. Candidates had their own test booklet and worked at their own pace for each question.	Video formats were slightly easier than the print formats. The print subsets were somewhat more reproducible. **DIFF**: When all of the subsets were compared, video cases were easier than the print cases. **DISC**: Case-total correlations were similar for the two formats within each type of study, although they tended to be low except for the arteriogram cases.		The equivalence of the video and print format in this study supports the use of the print format in national examinations. And due to the additional expense and testing time, video examinations should be used at the local level whenever resources are available. Correlations series with MCQ scores, candidate descriptors and experience found low to moderate correlations between MCQ and motion studies, and high correlations between print and video format. These studies showed a slight tendency to favour the video subset over the print. Results of the study propose that subsets of both formats are measuring some aspect of skill in reading motions studies. And that MCQ and motion studies might measure different knowledge or skills.

	Approximately four and half hours were allowed for each form[4].					
Anna, 1994 (72)	Two forms of CST[5] were administered in five centres to two different groups. The study explores the effect of interactive audio-visual (AV) format vs. text format on the examinees' decision-making process and problem-solving skill in a nursing competence computerized test (CST). The number of cases was 11 and participants were 263 nursing subjects.	To explore the effects on examinee performance in a CST exam in text versus audio-visual (AV) format and determine what impact this technology has on examinee performance.	86 subjects took Form I that consisted of nine text format cases followed by the two AV format cases (cases 6 and 10).	177 subjects took Form II that consisted of the same nine cases in the text format followed by two text format cases (cases 6 and 10). A total of 11 text cases in this group.	Internal consistency reliabilities for AV versus text format were similar to suggest that no reliability was lost by a change in format. AV formats had significantly lower mean scores than text format on (FRI)[6] items (lower mean scores indicate less risky behaviour). **DIFF**: No significant difference in examinee ability estimates expressed in (logit units) was seen between performance on both formats. **DISC**: Differences in ability estimates between AV and text formats were not statistically significant	Results suggest that examinees behave the same with regard to taking actions that benefit the patients. Results (longer time spent on AV format and lower FRI means) suggests that examinees exposed to AV format had more information or were more careful about a patient situation which points toward its fidelity to real-life experience.

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Lieberman, et. al, 2003 (70)	20 parallel test questions (video clips vs. text vignettes) of abnormal neurologic findings were given to 106 fourth-year medical students (who were randomly assigned) after their Neurology rotation through CBT. The two formats were equivalent in content.	Using patient video clips (as a testing method with higher fidelity) affects the psychometric properties of MCQs.	Identical to text questions except it was delivered using videos to present the findings.	Question with text vignettes describing physical examination findings.	Overall, video-based questions had similar difficulty and discrimination results compared to their equivalent text-based questions. Initial studies showed similar reliability with the text and video-based questions. There was a low-moderate correlation between video and text question scores.		Questions using both formats revealed similar result distributions of item difficulty and discrimination. The use of patient video-clips in CBT is feasible from a technical, practical and psychometric point of view. Further research is necessary to gather validity evidence of these types of questions. The low-moderate correlations between both formats suggest that the medium used (text or video) to present the information in the vignette might affect the nature of the assessed competencies within a knowledge domain.
					DIFF	**DISC**	
					Nonsignificant trends of lower difficulty values (i.e., more difficult) in videos compared to text	Nonsignificant trends of video format having a higher RPB (discrimination) than text format	

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Holtzman et al., 2009 (20)	USMLE introduced 43 Step-1 and 51 Step-2 unscored MCQs in two formats multimedia and text. Both formats were presentations of cardiac auscultation findings.	Use of multimedia (auscultation findings) has an impact on item characteristics (difficulty, discrimination and time)	Multimedia format (heart sounds and related videos of chest and neck vein movements) items placed at the end of a randomly selected section	Text format (auscultation findings using standard medical terminology) items were randomly inserted among other MCQs.	Results favoured text version of items for first-time examinees form the US and Canada, as well as international medical schools. Correlations between P values for multimedia and text version (In Step 1 and 2) were high. Similar correlations were found in item discriminations. Multimedia items required significantly more time to answer than the text version		Examinees are more able to interpret the described (text) version of auscultation findings than the more authentic multimedia. Audio items were constantly more difficult and less discriminating. Increasing fidelity and authenticity with multimedia format requires an increase in testing time (Multimedia items required on average, 30 to 60 seconds longer for a response than text versions) Using multimedia to present auscultation findings has a substantial impact on item parameters (difficulty, response time, to a more modest impact on item discrimination)
					DIFF	**DISC**	
					Multimedia items were significantly more difficult than the matched text version and required more testing time	Multimedia items were less discriminating than the text items for both groups in each step	

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Wendt & Harmes, 2009 (45)	Two fixed forms of 70 items were developed that contained both text and innovative items. (21 common items and 49 items unique to a test form). Each innovative item had a matched text version to make parallel test forms. Item position was the same for both forms, and innovative items were spread out across the test. 103 senior-level nursing students participated in 6 occasions (89 took the CBT exam and 14 were tested in individual	Examine the cognitive process required to answer different types of innovative items, as well as their statistical characteristics.	Different types of innovative items were created and refined. Items that were not possible to make a text version of appeared as innovative format in both test forms.	A text-based version of each innovative item was created and refined as much as possible. Rach pair measured identical content.	Difficulty values and item total correlation for both formats were generally similar. The video interaction items were generally more difficult in the innovative format. **DIFF**: Both formats were similar in difficulty. Innovative formats were more difficult When differences were noticeable. There were 2 text items that were more difficult. **DISC**: Both formats were generally similar in discrimination. Innovative formats were more discriminating when differences were noticeable. There were 5 text items with higher discrimination.		In general, innovative items were more difficult and more discriminating than paired-text items. Innovative items that were developed to test higher levels of cognitive processing did indeed and were rated higher based on the students' think-aloud.

	think-aloud sessions), out of the 89 students 42 took form A and 47 form B.					
Study	**Method**	**Hypothesis, question**	**Multimedia Item**	**Text Item**	**Results[1]**	**Conclusion**
Linjun Shen et al., 2010 (5)	Osteopathic Medical Licensing Examination developed 44 content-matched multimedia and text MCQs using heart sounds and videos. Items were matched in length, size, content, and wording as closely as possible.	Can multimedia items test additional elements of knowledge and skills (different constructs) from text-matched items if they perform differently? and secondly, how can we develop an effective and meaningful multimedia item?	Heart sound clips collected from daily practice was used. Video clips were short recordings of physical examinations and three cardiac images were used.	Text narration of the content of the parried-multimedia item was written. Additional information that was given in the MM format was described in the stem in the text format (the auscultatory site on the chest wall where the sound was recorded)	Paired Analysis: Nine pairs demonstrated significant differences in difficulty and /or discrimination. MM items were not uniformly easier for candidates across different ability levels, which may indicate that text and MM items test different elements of the same concept. MM items significantly needed longer time to answer by the examinees. **DIFF** / **DISC** Mixed results regarding difficulty level, MM items were slightly but significantly higher than text items / Discrimination was slightly lower for MM items; the difference was not significant.	When text narration was less direct, the MM made the items easier. However, MM made the item more difficult when it replaced the textbook terminology. Also, multimedia items seem to be measuring some construct that is different from what the text is measuring. Although difficult to explain, it does provide valuable information and might explain why differences in discrimination were more difficult to interpret.

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Vorstenbosch et. al., 2013 (54)	210 students were randomly assigned to 39 Extended matching questions (EMTQ) grouped in seven themes in the subject of gross anatomy. Most had multiple answers. Two versions were developed with identical items. Answer format was either an image or answer list in the MCQ response format (options). Altered response format per theme was used.	As stated in the article: (1) what is the influence of images as a response format on item difficulty and item discrimination in written assessments in anatomy? (2) What is the relationship between spatial ability and the influence of images in assessments? and (3) what is the influence of images as a response format on students' opinions about test items? [71]	Exam with an answer format (response format) contained a list of labelled images	Exam with an answer format (response format) contained an alphabetical list of textual options (that is an answer list)	Both examinations had similar overall difficulty and reliability. Variable effects from the images suggest a context-dependent interaction is taking place with item content. Cross-sectional images suggest an extra skill is being tested, while schematic data of foetal circulation suggests cueing.		The effect of images in assessment are divergent and not uniform (in their study results were presented by themes). Images influence item difficulty and to a lesser extent discrimination. Influence on scores is dependent on the type of image used that interacted with the content of the test item and had an effect on item difficulty and discrimination when used in the response format instead of an answer list. This may have implications for the validity of test items. Students with high spatial ability (SA) perform better in the exam as a whole and are less influenced by the form of the response format.
					DIFF	**DISC**	
					Some items had increased difficulty while others had decreased difficulty levels. Difficulty was more in cross-sectional images than answer list, less in schematic images and varied in other images.	Discrimination was more in cross-sectional images than answer list, less in schematic images and varied in other images.	
Study	**Method**	**Hypothesis,**	**Multimedia**	**Text Item**	**Results[1]**		**Conclusion**

		question	Item				
Holland J. et. al, 2015 (40)	Item analysis data from 3 consecutive years of six histology MCQs tests were used. A total of 195 items were categorized whether their stem (stimulus format) was purely textual (95) or included an image with it (100). The number of students per exam ranged from 277 to 347 1st year medical students with 60,850 student-item interactions. Questions tested only recognition and understanding	Does including images within the stimulus format of MCQs have a predictable or consistent influence on psychometric item properties?	Items with stimulus format (stem) containing an image named (illustrated text). Images used were relatively simple diagrammatic and histological images.	Item with stimulus format (stem) containing text (Text alone)	Overall, there was no influence on item difficulty, discrimination or point biserial that was seen **DIFF**: Items that included images had a higher mean difficulty however this difference was not significant **DISC**: Discrimination Index was unaffected by the addition of an image within the stem. Point biserial showed no differences either.		Image use in this context is statistically indiscriminating. It is suggested that if included within the stems it should be based on the principles of constructive alignment. Even with the advances of research in Cognitive Theory, and how visual and verbal material processing works, the evidence-base with respect to their effect in written examinations is scarce.

Study	Method	Hypothesis, question	Multimedia Item	Text Item	Results[1]		Conclusion
Notebaert. 2017 (49)	A retrospective analysis of 15 MCQ test items (undergraduate anatomy) given from two examinations was studied. Items were text-based and compared to items that were the same but had an image included in the stem (stimulus format).	As stated in the article: How does the inclusion of reference images affect the item statistics of anatomy MCQs, specifically the item difficulty and discrimination? (if item analyses significantly differed if items were presented as text-only or if they contained an anatomical reference image pertinent to the item.) (49)	Item with reference image appeared in an examination. Text and answer choices were identical in both presentations. Examination was given in paper format. (Some images were not critical to answering the item)	Text only item appeared in the examination. Text and answer choices were identical in both presentations. Examination was given in paper format.	There were some differences in item difficulty but these were not consistent to either text nor items with images.		Images do not significantly alter item statistics. It also does not indicate if images were helpful to students when answering the questions. Instructors should carefully select the appropriate image that would portray the correct information when selecting items to be included in an examination. Furthermore, item analysis should be closely examined to make sure there are no adverse effects.
					DIFF	**DISC**	
					Statistical analysis found limited, mixed differences in item difficulty	No significant differences for item discrimination were found for either of the item formats	

[1] Results are the overall general, details can be found in the original article.

[2] Fellowship of the Royal Australasian College of Surgeons

[3] Cardiovascular motion studies include (echocardiograms, ventriculograms, and arteriograms).

[4] Both forms (A &B) of echo cases contained 25 videos and 22 prints. The two forms (A &B) of ventriculograms/arteriograms had 12 videos of ventriculograms and 14 videos of arteriograms, and 9 or 10 print ventriculograms followed by 11 or 12 print arteriograms.

[5] Clinical simulation Test; [6] Flag, Risk Inappropriate items.

Appendix 3: Examples of proposed frameworks and dimensionalities for multimedia classification

Examples of proposed frameworks and dimensionalities for multimedia classification

Study Reference	Framework and Dimension Explained
Parshall, Harmes, Davey & Pashley, 2010	**1. Assessment Structure:** includes different ranges of formats (e.g., discrete items, item sets, constructed responses, situated tasks, simulated environments) **2. Complexity:** number and variety of elements of items the examinee must consider **3. Response action:** what physical interaction the examinee requires (e.g., type, click, drag-drop, record) **4. Media inclusion:** any variety of interactive media (i.e., audio, video, animations, graphics) in the response or stimulus format. **5. Level of interactivity:** how much the item reacts and responds with the examinee's input. Different degrees of interactivity may be presented ranging from a single-step item (make a selection), to limited feedback to examinees to a complex simulated patient interaction scenario. **6. Fidelity:** how much is the item's authenticity (i.e., degree of resemblance to real-life situation and context) **7. Scoring model:** relates to what type of response data and mode that are collected (e.g., recorded, the use of AI, multi-part items), and how examinees respond. This is then translated into quantitative scores.
Ruiz, Cook (38) (2009)	**1. 'Process visualized':** has three categories: Transformation, Translation, Transition. a) **Transformation:** A process that involves changes and alterations in key characteristics of the graphic form such as size, colour, shape, or texture. b) **Translation:** A process that involves positional changes (motion) from one location to another c) **Transition:** this process involves the appearance and disappearance of entities that change fully or partly. **2.** The domain of interactivity contains two categories interactive and non-interactive: 1. Interactive: where there is some degree of learner control over the animation sequence. 2. Non-interactive: where there is no learner control, the animation plays at a constant rate and time. **3.** Dimensionality can either be two or three-dimensional animation and finally, **4.** The level of abstractions can be divided into two categories: 1. Iconic, symbolic or representational: which means that the presented phenomenon is usually a realistic representation. 2. Conceptual or abstract form: which illustrates a non-visible concept.

Gulikers, Bastiaens (117)	1. **Task:** the task should reflect real-life situations and resemble the complexity and level, that means it should be relevant to what is perceived in the real world, and students need to be able to relate or link to the situation (117). 2. **Physical Context:** the safe and relaxed environment in which the assessment is being taken, the exam should reflect the way knowledge will be used in the clinical setting, time is also not reflected well between a quick response in an exam setting or more time to think maybe even over days in the real setting. 3. **3. Social Context:** difficult but should be able to resemble social processes of real life **4. Assessment Result or Form:** authentic results should be for performances that students can produce, permits making inferences about the underlying construct, has multiple indicators to make a fair conclusion, report the work **5. Criteria and Standards:** characteristics of assessment results that are valued. Should be explicit and transparent to the learner in advance, related to realistic performance and relevant competence.

Appendix 4: Combined view of Bennett et al. (1999) and Lui et al. (2001) for the development of multimedia in assessment

Combined view of Bennett et al. (1999) and Lui et al. (2001) for the development of multimedia in assessment

Category	Related Questions to Explore
Measurement and Purpose:	• What is being measured? (7) • Does it fit with the purpose of the construct? (7) • Does it provide relevant evidence? (7) • Does the benefit outweigh the harm when using the necessary technology for this? (7)
Content	• Should the multimedia and paper-based exam be the same? why? (80) • How can multimedia enhance the paper-based version? (80) • What abilities are provided by multimedia? (80) • What materials are needed to create a multimedia test? (80) • How many times can a test be exposed and reused after creating it? (80)

Resources/ Technical Issues:	• Test Development:
	  ○ Are the materials available? What are the resources for their availability? (7)
	  ○ What tools, resources, and methods are needed in the test development process? (7)
	  ○ How big is the multimedia bank? And what types and quality of media are available? (94).
	  ○ Regarding marking schema, is it feasible and how long will it take? (80)
	  ○ Is training or orientation towards multimedia use required? (7)
	• Delivery:
	  ○ What factors are needed to deliver multimedia to a wide audience? (80)
	  ○ What test centres are available? And what are the consequences of using them? (7)
	  ○ What available technical supports are present? (7)
	  ○ What equipment is needed from hardware and software etc. (80).
	  ○ What are the issues regarding security? What are the security measures that are needed? (80), (7)
	  ○ Is there a third party involved in the delivery and what are their roles? Or the delivery in house? (7)
	  ○ Is training required and who are the stakeholders involved? (7)
	  ○ What is needed to prepare examinees and orient them? (7)

Perspectives	• Students' Perspective:
	  ○ Students' perspective of multimedia testing (80)
	  ○ Feelings towards using multimedia: comfortable, anxiety, familiarity (80)
	  ○ Effect on students' learning: does MM help them learn the concept better? And in what way or How? (80)
	  ○ Are additional computer skills required when taking multimedia? (80)
	  ○ For those who do not have a computer, is taking MM fair to them? (80)
	• Faculties' Perspective:
	  ○ What do faculty think of mm use? (80)
	  ○ Feelings towards it: comfortable? Confusing? Complex? (80)
	  ○ What skills, knowledge, materials, equipment are required to deliver a test? (80)
	  ○ How long will it take to deliver a MM exam? Are faculty willing to spend more time on such tests? (80)
	  ○ What skills are required of them to create a multimedia exam? (80)

Financial categories	• What is the financial burden required to carry out this method? (7) • How much manpower is needed? Is it cheaper? Less manpower than paper-based? (80) • What are the cost for additional time, candidates, or methods when having a third party involved, with a pre-determined contract? (7) • Are the extra expenses worth the effort? (7) • Are the benefits worth the cost? (7)
Reliability	• Can it be reproducible and delivered on a large scale?" (7) • What is the technology of multimedia reliable enough for large scape testing? (80)
Outcome/Impact	• Can it be taught? and does it impact teaching and learning? (7) • What are the consequences of its use? Is it fair to certain groups? (7) • What advantages does CBT have over the paper-based test in regards to speed and grading? (80) • Is using multimedia powerful enough to realize its capabilities? (80) • Is this format worthwhile? Are the outcomes rewarding? (80)

Appendix 5: Types of validity

Types of validity

Types of validity	Definition
Criterion-Oriented Validity	Comprises of predictive and concurrent validity and is concerned with specific test criterion correlations. It is assessed by comparing test scores with an external variable (criteria) that is said to provide a direct measure of the behaviour in question (50, 154) For example: Does this IQ test predict study results? The study results are the criterion in this example. (50)
Predictive Validity	The extent to which a person's future level on a criterion can be predicted from previous test performance. (e.g., a candidate's job performance after graduating; or SAT scores predict GPA scores in college. Attempts to answer the question: do scores predict future performance? (1, 42, 153, 154)
Concurrent Validity	The extent to which the test scores estimate an individual's present standing on the criterion For example, when a multiple-choice form of spelling test is substituted for taking dictation (1, 42, 153, 154).
Content Validity	How well the test content samples subject matter that one draws conclusions from. It relates to what the blueprint covers (1, 153, 154). For Example: Do items on a particular IQ test cover all aspects of intelligence? (50)
Construct Validity	Tries to measures a trait, quality or attribute that is not clearly formulated or defined, and is interpreted by a test. Examples of constructs are intelligence, personality, attitude, creativity measurement, professionalism, teamwork, diagnostic reasoning, cognitive function. Construct validity includes almost all forms of validity (1, 50, 153, 154)For example: Does a particular IQ test measure intelligence? (50)
Face Validity (acceptability)	Ensures that scale items are actually measuring what they set out to measure. In other words, does the test appear to the examinee as it should be, is it testing what it is supposed to test? (1). For example, a high-fidelity patient simulation exam would be used to assess certain clinical/surgical performances.
Consequential Validity	It deals with psychological, social, intended and unintended consequences that arise from the use of the test (159 p.176). For example: pass/fail because a test also measures something else other than the construct (i.e., CIV).
Reliability	Reproducibility, and tries to answer: Will we get the same results consistently? (1)
Feasibility	How practical is it, from a fiscal, logistical, technological and staffing point of view? (1)

Appendix 6: Outline of the *Standards* from AREA, APA and NCME (2014)

\

Outline of the *Standards* adapted from AREA, APA and NCME (2014) (155)

Standard	Overarching Standard (relates to)
Standard 1.0 (Validity)	Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.
Standard 2.0 (Reliability/Precision)	Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.
Standard 3.0 (Fairness)	All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimise construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinee in the intended population.
Standard 4.0 (Test Design and Development)	Test and testing programs should be designed and developed in a way that supports the validity of interpretation of test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.
Standard 5.0 (Scores and Scales)	Test scores should be derived in a way that supports the interpretation of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.
Standard 6.0 (Test Administration, Scoring, Reporting and Interpretation)	To support useful interpretations of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected.
Standard 7.0 (Supporting Documentation for Tests)	Information relating to tests should be clearly documented so that those who use test can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores.
Standard 8.0 (Test Takers' Rights and Responsibilities)	Test takers have the right to adequate information to help them properly prepare for a test so that the test results accurately reflect their standing on the construct being assessed and lead to fair and accurate score

	interpretations. They also have the right to protection of their personally identifiable score results from unauthorized access, use, or disclosure. Further, test takers have the responsibility to represent themselves accurately in the testing process and to respect copyright in test materials.
Standard 9.0 (Test Users' Rights and Responsibilities)	Test users and are responsible for knowing the validity evidence in support of the intended interpretations of scores on tests that they use, from test selection through the use of scores, as well as common positive and negative consequences of test use. Test users also have a legal and ethical responsibility to protect the security of test content and the privacy of test-takers and should provide pertinent and timely information to test takers and other test users with whom they share test scores.
Standard 10.0 -13 (Testing Applications)	Standards 10.0, 11.0, 12.0 and 13.0 have to do with testing applications of certain tests, namely psychological testing and assessment, workplace testing and assessment, educational testing and assessment and tests for program evaluation, policy studies and accountability respectively.

Appendix 7: Examples of validity frameworks

Examples of validity frameworks (164)(159 p.171-173)

Frameworks	**Inferences**				
Downing Based on *the Standards*	**Content (BP, domain, item)**	**Response Process**	**Internal Structure**	**Relationship to other variables**	**Consequences**
	• Test BP • BP representation of the domain • Test specification • Item content matches test specification • Representativeness of items to domains • Relationship of test content to the domain • Quality of items • Qualification of item writers • Sensitivity review	• The familiarity of students to format • Quality control for scoring • Key validation of initial scores • Accuracy of Combining scores from different formats • Quality control of final scores • Subscore analysis • Accuracy of pass/fail decisions to scores • Quality control of score reporting • Relaying score interpretation in an accurate understandable way to students	Parameters of Item analysis (DIFF, DI, item characteristic curve, TCCs, inter-item & item-total correlation) Score reliability SEM Generalizability Dimensionality Item factor Analysis Differential item functioning Psychometric modelling	Correlation with other relevant variables correlations with similar tests correlations with dissimilar measures Test-criterion correlation) generalizability of evidence	Impact of scores on student and society The consequence on students future learning Positive consequences outweigh negative consequences Methods used for pass/fail scoring are reasonable. Pass/fail consequences False positives/negative of passing or failing students Instructional/learner consequences

Kane (157, 166)	Scoring	Generalization	Extrapolation	Implication
	Translating an observation into one or more scores -Need expert judgment, scoring rubric/criteria, audits, form standardisation and equating, reliability analysis, rater selection and training,	Using scores as a reflection of performance in a test setting Generalizability, reliability study, IRT, sampling strategy and size.	Using scores as a reflection of real-world performance Examine relationships between scores and criteria, the scope of the test, Authenticity of assessment context	'applying the score[s] to inform a decision or action' Pass/fail standard (e.g., Angoff method), Effectiveness of actions based on results, intended or unintended consequences of testing.
	Evidence are collected to support each of these inferences			

Appendix 8: The 12 Components (steps) for an effective test development process

Appendix 8, containing the table **The 12 Components (steps) for an effective test development process** adopted and adapted from Downing & Haladyna From the Book Handbook of Test Development by S. Lane, M. Raymond & T. Haladyna **(165, 168 p.4-5)** has been removed due to Copyright restrictions.

Appendix 9: Strengths and weakness of mixed method research

Strengths and weakness of mixed method research

Strength	Weakness
The use of multiple theories, perspective and methods adds strength to the educational research and adds insights that may be missed using a single method	May be difficult for the researcher to carry out both aspects of the study and, therefore, may require a research team
Words, pictures and narrative can be used to add meaning to numbers and numbers can be used to give precisions to pictures and words.	The researcher is required to understand and know about different methods and how to use them and mix them properly
Qualitative data can help identify problems in quantitative data, provide feedback and help correct them	It is more expensive and time-consuming
Quantitative data can add amount and frequency to qualitative data	Some aspects of mixed-method research such as how data are integrated and analyzed how to deal with conflicting results and the problems of paradigm mixing are still being worked out
Statistics can be better understood from the insight of the qualitative part (variation in human characteristics)	Researchers who believe to stick with one paradigm tend to contend the mixed paradigm (159 p.473)
Can obtain a fuller and deeper answer to research questions	
More than one method can be used and, therefore, the strength of one method can overcome the weakness of another	
Can provide validation through a better convergence of findings	
Multiple stakeholder involvements add to the understanding of their realities	

Appendix 10: Principles of questionnaire construction

Appendix 10 containing the table of **Principles of questionnaire construction adapted from Johnson and Christensen ((159) p.193)** has been removed die to Copyright restrictions.

Adopted from Johnson RB, Christensen L. Educational Research: Quantitative, Qualitative, and Mixed Approaches. 2016. (159 p.193),

Appendix 11: Study questionnaire

Questionnaire

Thank you for taking the time to fill out our questionnaire. Your feedback is a vital part of the development process and improvement of our exams and we are looking forward to receiving your completed form.

1. **PLEASE <u>CHECK</u> THE BOX YOU AGREE WITH THE MOST**

	1. Strongly Agree	2.Agree	3.Disagree	4. Strongly Disagree
• Exams Instructions were clear and appropriate	☐	☐	☐	☐
• Exam questions provided good coverage of the curriculum	☐	☐	☐	☐
• Exam covers materials on which I expect to be tested on	☐	☐	☐	☐
• The exam reflects important relevant topics from my daily practice	☐	☐	☐	☐
• The layout of the questions was easy to follow	☐	☐	☐	☐
• The language used in the exam was understandable	☐	☐	☐	☐
• There was enough time to complete all the questions	☐	☐	☐	☐
• Exam requires me to do more than recall facts	☐	☐	☐	☐

2. Regarding Multimedia Questions **PLEASE <u>CHECK</u> THE BOX YOU AGREE WITH THE MOST (section A, B, C):**

(2.A) Regarding Multimedia (MM) Questions (Images/Videos) :	**1. Strongly Agree**	**2.Agree**	**3.Disagree**	**4. Strongly Disagree**
• Multimedia questions felt more realistic and reflect what is emphasized in the clinical practice/real-life situations	☐	☐	☐	☐
• Multimedia questions have more information about the patient's situation	☐	☐	☐	☐
• Multimedia questions were more relevant to my clinical practice	☐	☐	☐	☐
• The multimedia questions challenged me to do original thinking	☐	☐	☐	☐
• The multimedia questions made me feel more involved and able to understand the patient's problems better	☐	☐	☐	☐
• The multimedia questions made decision making more certain because it offered more information than words alone could convey	☐	☐	☐	☐
Multimedia questions were easier to answer	☐	☐	☐	☐
• The use of multimedia complements/completes the presentation of the problem being presented and tested (makes it better)	☐	☐	☐	☐

3. **Regarding Questions WITH Image (IF NOT APPLICABLE PLEASE GO TO Q3)**

(2.B) Regarding Questions WITH Images	**1. Strongly Agree**	**2.Agree**	**3.Disagree**	**4. Strongly Disagree**
• The images used were clear	☐	☐	☐	☐
• Image enlargement function was useful	☐	☐	☐	☐
• The use of images in the questions is good	☐	☐	☐	☐
• The use of images added strength to the exam	☐	☐	☐	☐
• The use of images contributes to my clinical work	☐	☐	☐	☐
• The use of images stimulated me to think	☐	☐	☐	☐
• The use of images challenged my clinical reasoning skills	☐	☐	☐	☐
• The use of images required me to do more than recall facts	☐	☐	☐	☐
• The use of images resembled medical conditions similar to those actually encountered in the work setting	☐	☐	☐	☐

4. Regarding Questions WITH Videos (IF NOT APPLICABLE PLEASE GO TO Q3)

(2.C) Regarding Videos (If NOT APPLICABLE please go to the next Question "Q3")	**1. Strongly Agree**	**2.Agree**	**3.Disagree**	**4. Strongly Disagree**
• The videos used were clear	☐	☐	☐	☐
• Video enlargement function was useful	☐	☐	☐	☐
• The use of videos in the questions is good	☐	☐	☐	☐
• The use of videos added strength to the exam	☐	☐	☐	☐
• The use of videos contributes to my clinical work	☐	☐	☐	☐
• The use of videos stimulated me to think	☐	☐	☐	☐
• The use of videos challenged my clinical reasoning skills	☐	☐	☐	☐
• The use of videos required me to do more than recall facts	☐	☐	☐	☐
• The use of videos resembled medical conditions similar to those actually encountered in the work setting	☐	☐	☐	☐

5. Regarding Computer-Based Testing (CBT) PLEASE CHECK THE BOX YOU AGREE WITH THE MOST

3. Regarding Computer-Based Testing (CBT):	**1. Strongly Agree**	**2.Agree**	**3.Disagree**	**4. Strongly Disagree**
• This experience with CBT was better than expected	☐	☐	☐	☐
• the orientation video was helpful in navigating the exam	☐	☐	☐	☐
• The navigation options (scroll, mark, review, etc.) were pretty easy to use	☐	☐	☐	☐
• The “Review function” makes it easier to go back and review the test compared to paper-based exams	☐	☐	☐	☐
• The “Calculator function” was sufficient to do simple math	☐	☐	☐	☐
• The “Scroll function” was distracting (e.g., cannot see both the question and image at the same time)	☐	☐	☐	☐
• The computer was fast enough that it does not take long for questions to appear on the screen	☐	☐	☐	☐
• The font size was appropriate	☐	☐	☐	☐
• I felt comfortable reading from the monitor screen	☐	☐	☐	☐
• The computer exam was easy to navigate	☐	☐	☐	☐
• I prefer CBT exams over paper-based exams	☐	☐	☐	☐

Overall Remarks: PLEASE CHECK THE BOX YOU AGREE WITH THE MOST	**1. Strongly Agree**	**2.Agree**	**3.Disagree**	**4. Strongly Disagree**
• The staff were professional	☐	☐	☐	☐
• Being in a cubicle makes it easier to concentrate	☐	☐	☐	☐
• An extra paper and pencil could be useful/helpful to take notes	☐	☐	☐	☐
• I was satisfied with my overall experience with the exam	☐	☐	☐	☐
• Temperature of the room was comfortable	☐	☐	☐	☐

5. Did you experience any technical problems during the exam? ☐Yes ☐No

- If yes, please explain

- Overall comments about the exam

THANK YOU FOR TAKING THE TIME TO COMPLETE THE QUESTIONNAIRE
GOOD LUCK

Appendix 12: Information sheet and consent form

Information Sheet
Exploring the validity of multimedia written assessment in Saudi Arabia

You are invited to take part in a research study that is taking place from October 2012 until October 2015. You will receive information about this project that aims to explore the validity of using multimedia multiple-choice questions (MCQs) in written examinations.

Purpose of the study:

Help raise the quality of practising doctors by raising the assessment standards of safety and clinical competence in high-stake written examinations. MCQs can be written to test higher cognitive skills such as analyses, interpretation and evaluation if structured correctly and supported with the innovative technology and the use of computer-based testing. This can be aided by using a set of systematic methods and recourses to help validate not only the examination questions but also the whole test development and assessment processes.

Participants of the study:

I am inviting residents of all levels who are in the speciality of Emergency Medicine. I am interested in the views, perspectives and experiences of Emergency medicine physicians who have been through SCFHS written examinations and can relate exam contents and relevancy to their daily life practice in the emergency care setting. I am also interested in your perspectives and feedback on some of the steps taken to construct the examination process.

What would be involved?

I would like you to take part in a focus group discussion. I will invite 8-10 residents to the group, some of whom you might know already as your colleagues. There will be a facilitator who will ask questions and facilitate the group discussion, as well as two note-takers to write down the ideas and what is expressed in the group. We would like your feedback on how you felt that you performed on the examination, how you perceived the examination to be in terms of difficulty, relevancy to your speciality and how well the multimedia items resembled real-life cases. The focus group will be arranged in advance (approximately 2-3 months beforehand) at the most convenient time for participants and will take around 1 hour and 30 minutes. If you have any questions or concerns regarding the study you are welcome to contact Thuraya Kattan (contact details below). The discussions will be tape-recorded so that I might have a record of what was said, this will help me to further understand and analyse what was said in the group.

If you are not comfortable to participate in front of the focus group and feel more comfortable to share your views in an individual meeting, that could be arranged. Terms of privacy and confidentiality will be the same as that of the focus group (please see below).

What will be done with the information gathered?

The discussion of the focus group will be transcribed into print and will then be analysed by myself and the Plymouth University researchers. All participants will be given the researcher's draft analysis report of the focus group to read and if you wish to give your written or verbal comments on it. The transcript will only be used for the purpose of the study and will not be used for any other purpose. It will also be dealt with by the primary investigator and the Plymouth University researchers and no one else. The information gathered from the discussions will be a source for my PhD thesis. The results of this study and some of the transcripts might be used for writing up and publishing articles in academic journals or may be presented in meetings or posters sessions, however, your identity will be protected (confidentiality and anonymity will be protected, see below). If you are interested, you are welcome to have a copy of the final transcription, its analysis, results and of the articles before they are published if you wish to receive them.

Confidentiality and Privacy:

Your participation in this research is voluntary, and you are free to refuse to participate in this study or withdraw at any time without giving reasons and without your decision affecting your future career or training as a resident/specialist. Participation involves attending the focus group discussion and contributing as much or as little as you please. Everyone attending the focus group will be asked that the discussion be confidential, to respect the privacy of others and not to disclose any information outside the focus group. All names will be changed and kept anonymous in the transcript, and confidentiality on specific answers will be assured so as no one can be identified from the transcript or linked to any results. Your place of work or level of training will not be mentioned nor appear in the printed copy. Only the researchers involved in the study will read your responses. The transcript will be securely stored and kept in a locked place, and all electronic materials will be encrypted and password protected and access to files will be restricted. Audiotapes will privately be demolished within five years of completion of the research. You will not be identified in any report or presentation arising from the study. However, feedback concerning the overall outcome of the study will be offered as an institutionalized report. If you agree to participate in this study, then responding to this invitation is required, and you will be asked to sign a consent form.

Benefits, Risk Estimation and Protection from Harm:

The study involves you taking part in a focus group discussion to share your views and perspectives on the written examination and its construction process. Your participation may benefit yourself, as well as other specialities in improving the examination process and raising the quality and standards of the assessment process. This study does not pose any physical risk or harm to the participants apart from the usual discomfort of a conversation experience. The focus group discussion is not marked and will not contribute to mark or evaluation. It may be possible that you experience some stress or frustration depending on what is discussed in the group, but you are free to leave at any time if you wish. The researcher or a resource from the Commission will be available for support and to privately discuss any discomfort with you afterwards if you wish. Participation in this study will not affect your current nor future exam results or training evaluation. Participants will be anonymous and will not be identifiable by the information given. The information drawn from this focus group discussion is for research purposes only.

If you have any further questions or concerns about the study, you can contact the principal investigator

This research has been approved by the Saudi Commission for Health Specialities, Plymouth University Peninsula School of Medicine and Dentistry.

Consent Form:
Exploring the validity of multimedia written assessment in Saudi Arabia

The purpose of this study is to raise the quality of practising doctors by raising the assessment standards of safety and clinical competence in high-stake written examinations. MCQs can be written to test higher cognitive skills such as analyses, interpretation and evaluation if structured correctly and supported with the innovative technology and the use of computer-based testing. This can be aided by using a set of systematic methods and recourses to help validate not only the examination questions but also the whole test development and assessment processes.

☐ I have read all that has been mentioned in the information sheet and understand all that has been mentioned in the above study.

☐ I have had the chance to reflect onto the information given, ask questions to clarify points, and have received adequate answers to them.

☐ I recognize that my participation in this study is voluntary and that I am able to withdraw whenever I want at any point in time of this study without giving a reason and without my decision affecting on my future career or training as a resident.

☐ I understand and agree that my words might be quoted directly, confidentially and without identifying who I am. I understand that my name will be kept anonymous, and a made-up name may be used instead.

☐ Audio-tapes will be stored in a secured place.

☐ I understand and agree that the researcher may publish the results of the discussion and it may contain some quotations made by me anonymously.

By ticking all the above boxes and signing this consent form you are showing that you completely understand the above information motioned and agree to participate in the focus group discussion.

I agree to contribute as a participant in this study.

Participant's Name:
Signature:
Date:

If you are not comfortable to participate in front of the focus group and feel more comfortable to share your views in an individual meeting, that could be arranged. Terms of privacy and confidentiality will be the same as that of the focus group (please see below).

Two copies will be made of this consent form, one for the participant, and one for the researcher.

_____	_____	_____
Signature	Signature	Signature
_____	_____	_____
Name	Name	Name
_____	_____	_____
Position	Position	Position

Appendix 13: Points checked by specialist and content expert reviewers for items

Points checked by specialist and content expert reviewers for items

Component	Points Checked by Specialist Reviewer	Points checked by Content Expert Reviewer
Stem	• The stem was reviewed for the inclusion of relevant information, defining abbreviations • The scenario was checked for the appropriate sequence of clinical presentation, rephrasing of the question • All vital signs, lab results were checked that they were in the appropriate format (e.g., bold, SI unit, space, font and size)	• The stem was reviewed for accuracy of content and that the given information was sufficient. • That it was a realistic case, commonly seen in students' daily practice, clear, no doubt about the given information. • Elimination of any names, stereotypes • Describes changes that occur and not interpreting the findings, gives normal values when data are used.
Question Line	• The question was reviewed that it was clear and closed-ended. • That no unnecessary or additional information was included in the questions. This would be removed and inserted in the stem.	• Appropriate – (clear, specific, asks in relation to scenario)
Options	• Balancing of the options and randomization of the correct answer. • Options are arranged chronologically or numerically • Options did not contain clues or IWF	• Key – was checked for appropriateness, correctness and that there was no more than one correct answer. • Distractors – was checked for plausibility, and that there were no fillers. • That all options were homogenous and similar to key answer in length and complexity
Item as a whole	• Editorial review: clarity, grammar, spelling, punctuation and capitalization error (27 p.228) • Checked for any IWFs against item writing guidelines, which was then removed and edited • Check for any grammatical, spelling errors and corrected accordingly • Review for any possible cues or clues that might lead to the correct answer • Sensitivity and fairness review: check for insensitive content, language or stereotyping of person, offensive content that implies racial, ethnic and other groups. • **Special language care review:** ○ Review that appropriate wordings were used and that and lab data used as were unified throughout the items e.g., (Na Vs. Sodium), (oxygen saturation or SpO_2), using	• Checked for appropriateness – (aligned with Test Blueprint and level) • Checked for authenticity • The content was checked that it was up to date with the medical literature. • Check for any ambiguities or difficulties in the question, flaws and level of question • Cognitive demands review • Content review (content classification)

	man, woman, male, female, boy, girl (<19 years old), unifying all MVA to MVC). ○ Review for the presence of any unfamiliar words (e.g., guiac was replaced by stool +ve for occult blood)	
Images, videos and their descriptions:	• Ensure that no description is present in the scenario that would duplicate the image • Ensure that images used are consistent with the given scenario (e.g., age, gender, site, side: left/right) • Ensure that the wordings “see clinical image”, “see X-ray”, “see scan” video was included in items associated with multimedia to notify candidates of its presence, in case it did not appear on the exam screen. • All images were edited. This included checking format, numbering them, putting them in a separate folder, remove any names or MRN number, labelling A/B on images if there were more than one image per question.	• Ensure that MM and images were copyrighted. • That MM and image were clear and relevant to the scenario • That image added to the level of the question.
Other	• Identifying questions with calculators • Checking the appropriateness of classification and level of question assigned	• Reference – (up to date) • Abbreviations – (clear, list, appropriate)

Appendix 14: CBT exam specification

Computer-based test exam specification:

A) General exam specification:

- The exact date of the Examination needed to be scheduled in order to reserve seats in the computer labs to administer the test.
- The exam bank would contain 160 questions and would be delivered into two forms
- A unique master code was generated for all exam items, as well as all the multimedia by the test administration industry (this was delivered in a form of a template).
- A unique registration code was developed by the Commission for each resident.
- Test administrator needed to match each computer with residents' new ID code, as well as with the specified exam form, they would receive.

B) Form specification:

- The exam comprised of two forms each with a unique label to identify which contained the multimedia items or the texted matched items.
- The first form (Form-A) would contain 130 questions (100 original promotion questions) in addition to 30 text-based beta questions. The second form (Form-B) would also contain 130 questions (the same original 100 promotion questions) in addition to 30 multimedia beta questions.
- The first one-hundred questions would appear in both forms A and B in the same order provided by the Commission. (i.e., questions 101-130 would appear in form A only in the same order provided, and questions 131-160 would appear in form B only in the same order provided).
- Questions will follow the same appearance and order as delivered by the Commission in their documents.
- Characteristics of question layout in each form was provided by the Commission using the given item and multimedia master code and then linked to the residents' registration master code.

C) Test Blueprint and Classification:

- Exam blueprint for both 100 promotion questions, as well as the 30 paired questions needed to be delivered to the test delivery service in advance.
- Classification under each question, as well as the blueprint details needed to be submitted.
- All the questions in the blueprint supplied were used in this examination as this was a one-time examination.
- Two new classification sub-heading was added these were "Forms" and "mastercode" and were included under each question to identify on which form (A or B) the item needed to appear on and what its unique code was.
-

D) Multimedia specification:

- Multimedia questions used in the exam comprised of images and videos and needed to be submitted at an early stage for quality testing.
- Multimedia given within this examination was not permitted to be used in any other examination or to be stored in the test administrator's banks.

E) Logistic specification:

- Reservation of the computer test centres/labs in three regions of Saudi Arabia (Central, Eastern, and Western) needed to be arranged and confirmed at an early stage, this required setting the exam date and time with the EM scientific committee.
- An orientation video of the test administrator's software needed to be demonstrated to residents before the examination to orient them on how to navigate computer-based tests.
- A paper-based survey needed to be developed in order to be distributed at the end of the examination to candidates. The feedback would be used to make the improvements and adjustments for the main project the following year.
- Detailed exam specification and test setup including images, layout, computer screen, etc. were covered through the completion of a test specification document (TSD) Appendix 17.
- For quality assurance, questions would be uploaded through a secure test portal (STP) system
- The exam time needed to be increased from two to three hours to compensate for the increase in number in the original exam items, as well as the addition of the 30 paired questions
- The paired questions were not included in the candidate's final marking, however psychometric properties were needed

F) Scores and reports:

- Residents' results were requested to not appear on the screen at the end of the examination (review process according to item statistics)
- Item analysis reports on the performance of the questions were requested to be available for all items marked and unmarked.
- Result report was requested to be sent immediately to the SCFHS STP system for psychometric analysis to supply the exam committee with the results.

Appendix 15: Strength and weaknesses of MCQ

Strengths and weaknesses of MCQs

Strength of MCQs
Can measure the higher level of thinking (196, 233, 234)(27 p.66)
Test a wide range of content domain and learning objectives (21, 98, 187, 196, 200, 234, 235) (2, 27 p.65)
Valued by test developers for their higher reliability (27, 151 p.65)(233, 234)
Easy to mark (27, 151 p.66) and can be scored automatically (196, 233, 234)(27 p.66)(165 p.289)
Is characterized by being objective in scoring (the same results will come out with different content experts, agreement on the same key answer) (187, 200, 233) (27 p.66)(165 p.289)
Suitable for large-scale testing (2, 151, 196)
Validity evidence for selected-response format is strong 66 (165 p.289)
Have many validity advantages in measuring cognitive ability and achievement (165 p.289) through establishing content validity by allowing a representative sampling of the cognitive domain. It also decreases the threat to validity by avoiding construct underrepresentation
Familiar to learners and most feasible to use in schools and universities (151)
Can discriminate between high and low ability students if properly constructed. (187)
Robust psychometric properties (236)
Well-constructed items are able to test higher levels of cognition (reasoning) and can discriminate accurately between high and low ability students (2)(27 p.38)
Has a greater efficiency than other test formats because it is cost-effective, takes less time to score and good items are reusable for future exams reducing test preparation time for the future (27 p.66), (165 p.289)
Well-constructed items developed by content experts and are edited, reviewed, administered and score are defensible (165 p.289)
Takes less time to answer compared to constructed-response format (235)
Weakness of MCQs
Difficult to construct successfully (2, 151, 185, 234) and Cannot measure psychomotor skills or production skills (165 p.289)
Prone to cheating (151)
Time-consuming (185, 187, 200)
Success depends on the appropriateness of the distracters (185)
Often test recognition of knowledge over higher cognitive process (151, 187) and difficult to test diagnostic ability (234).
Critics of MCQ characterize it by being artificial by providing examines with a predefined list of possible answers (165 p.289)
Has a clueing effect, that may aid in finding the correct answer (165 p.289)
Poorly written items with flaws are a major source of CIV (165 p.289)
Not authentic, is artificial, not a reflection of real-life clinical situations (151, 233)
Difficult to identify a student's learning needs (no information on the student's thought process to the answers is provided).(235).
Allows for guessing which affects the marking process (185, 187, 234).
Trivial content and poorly trained item writers introduce flaws to the items and some see this as a weakness (165 p.289)
Guessing the answer is viewed by some to be a major weakness (165 p.289)
They inhibit students from creative expression and original thinking (185)
Lacks face validity as it doesn't reflect clinical practice (234)

Appendix 16: List of possible MM-TXT topics

List of possible MM-TXT topics

Specialty	Diagnosis	Description	MM Type
Cardiology	Tamponade or Pneumothorax	Case with neck distention (internal jugular distention) with tamponade or pneumothorax	**Image or Video**
	Ventricular Tachycardia	Normal Rhythm goes to V. Tach + fusion beat	**Video**
	Pericardial Tamponade	Echocardiogram with tamponade	**US (ECHO)**
	Pulmonary Embolism	Echocardiogram, ventricular collapse (or IVC collapse on US)	**US (ECHO)**
Gastroenterology	Pancreatitis	Edema around the pancreas	**CT**
	Spontaneous bacterial peritonitis	Procedure of ascetic tap with dark coloured fluid drainage (infected) Or wrong landmark	**Video**
	Volvulus	Radiological Sign of volvulus	**X-ray**
	Pseudovalvulus	Psuedo-vovulus (psuedo-obstruction)	**X-ray**
	Nasogastric tube	NG tube in place or too high	**X-ray**
	Calculous cholecystitis	Impaction in CBD (US with stone in GB)	**US (US with stone in GB)**
	Appendicitis	Sonographic sign of +ve appendicitis	**US**
Infectious Diseases	Pneumonia	PCP	**X-ray**
	Measles	Measles	**Picture**
	Diphtheria	Throat: diphtheria	**Picture**
	Diabetic foot (Cellulitis)	Diabetic foot and subcutaneous emphysema	**Pic / X-ray: preferably both**
	Bilateral facial palsy	Bilateral facial palsy or unilateral if not available	**Video/Image**
Neurology	Seizure	Seizure, starts as partial seizure and progresses to generalized	**Video**
	Pseudoseizure	Pseudoseizure, patient pretending to have seizure	**Video**
	Any available diagnoses	Abnormal Reflexes	**Video**
	Myasthenia gravis	Myasthenia gravis	**Video**
	Cerebellar lesion	Cerebellar sign	**Video**
OB/GYN	Pre-eclampsia	Pregnancy with pre-eclampsia symptoms (image of edema of face or hand)	**Picture/Video**:
	Mastitis	Mastitis in a breastfeeding mother	**Picture**
	Deceleration	CTG showing Deceleration	**CTG Rhythm**
	Premature rupture of the membrane	+ve Test for premature rupture of the membrane	**Picture**
	Any type of abortion	abortion	US
	Ectopic pregnancy	Ectopic pregnancy	US
Renal	Urinary Bladder injury	Cystogram	**X-ray/Video**

	Stone in urinary tract	stone	**CT or X-ray** with stone
	testicular torsion	testicular torsion	**Picture**
	Stone in the urinary tract	stone	**US or CT with stone**
	Pyelonephritis/ renal abscess	Perinephric edema or abscess or bleeding	**US**
Orthopaedic	Less franc fracture	Less franc fracture	**X-ray**
	Tendon injury	Tendon examination in hand with an abnormal finding	**Video**
	Calcaneus fracture	Calcaneus fracture	**Picture or X-ray**
	Shoulder dislocation	Shoulder dislocation/or relocation	**X-ray**
	Boxer fracture	Boxer fracture	**X-ray/Pic**
	Tenosynovitis	Tenosynovitis (image and test)	**Video or pic**
	Paronychia	Paronychia	**Picture**
	Phelon	Phelon	**Picture**
Paediatrics	Infantile spasm	Infantile spasm	**Video**
	Child abuse	Child abuse, e.g., abuse burns	**Picture/X-ray**
	Tooth fracture (avulsion)	Pediatric Tooth fracture (avulsion)	**Picture**
	Anaphylaxis	Hereditary angioedema	**Picture**
	Orbital cellulitis	Orbital cellulitis	**Picture**
	Periorbital cellulitis	Periorbital cellulitis	**Picture**
	Diaphragmatic hernia	Diaphragmatic hernia (neonates)	**X-ray**
	Foreign body	Foreign body like a disc battery or coin in the oesophagus or abdomen	**x-ray**
	Foreign body	Foreign body in the chest causing collapsed or hyperinflated lung	**X-ray**
	Neonatal jaundice	Neonatal jaundice	**Pic**
	Fingertip Amputation	Fingertip Amputation	**Pic**
	Tufts fracture	Tufts fracture	**X-ray**
Toxicology	Tricyclic antidepressant overdose	TCA ECG strip	**ECG**
	Digoxin effect or toxicity	ECG changes of digoxin	**ECGs**
	Body packer	Body packers	**X-ray**
EMS	Scenario with two cases	Scenario with two cases to triage in a disaster	**Video**
Neurology	Abnormal Light reflex	Pupil reaction, location of the problem	**Video**
	Neurofibromatosis	Neurofibromatosis	**Picture**
	Meningitis/encephalitis	LP abnormal fluid (dark, bloody)	**Video or Pic**
ENT	Foreign Body in ear	Foreign Body in ear	**Picture**
	peritonsillar abscess	Throat: peritonsillar abscess or similar	**Picture**
	Otitis media or if not available, Normal tympanic membrane	Audioscope (pressure) of the tympanic membrane to show mobility	**Video**
	Nasal foreign body	Nasal foreign body	**Picture**
	Epistaxis	Nasal packing	**Picture**

	Epiglottitis	Epiglottitis	**X-ray**
	Thyroid masses	Thyroid masses	**Picture or US**
	Vascularity of the nose	Epistaxis (area of bleeding) diagram	**Drawing**
	Rupture tympanic membrane	Rupture tympanic membrane	**Picture**
	Sub mental abscess/edema	Sub mental abscess/edema	**Picture**
	Glaucoma	Mid dilated pupil and red eye	**Picture**
Toxicology	Methemoglobinemia	Methemoglobinemia (chocolate blood colour) or video of blood extraction	**Picture/video**
	CO poisoning	CO poisoning, redness in skin (flushed)	**Picture**
	Opioid overdose	Pupil pinpoint (opioid)	**Picture/video**
	Stimulant overdose	E.g., amphetamines causing Diapheretic skin and tachycardia & irritability.	**Picture/video**
	Ethylene glycol container	Ethylene glycol (anti-freeze)	**Picture**
Trauma	Stab wound	Stab wound/ penetrating injury insito in a stable patient	**Picture**
	Chest Stab Wound	Stab wound in the chest	**Picture**
	Subungual Hematoma	Minor trauma (nail avulsions) or hematoma under nail	**Picture**
	Free fluid in the abdomen	FAST bleeding	**FAST**
	Subconjunctival haemorrhage	Subconjunctival haemorrhage	**Picture**
Resuscitation	Sinus Arrhythmia	Waves of monitor breath to breath variation	**Video**
	Procedure	Intubation procedure (vocal cord not moving)	**Video**
	Facial Trauma	Facial trauma in a bearded person with a short neck (i.e., difficult airway)	**Pic or Video**
	Procedure	Cricothyroid for the landmark of cricothyroidotomy	**Pic**
	Electrical Alternans	Video/Pic of a monitor or Rhythm Strip: electrical alternans	**Video/Pic** of a monitor **OR Rhythm Strip**
	Pneumothorax	Pneumothorax	**X-ray**

Appendix 17. Test specification document (TSD)

Test Specification Document (91, 155, 168)		
Sections	**Brief explanation of section content**	**Examples**
1. General Information	Test design specialist, client service manager, test developer, psychometrician's involved, Exam name abbreviation, title & type of exam	Names, the contact information of client, manager, etc. Type of exam beta exam.
2. Exam Approvers	Client and Test design specialists	Contact information and roles
3. Project Scope	1. Type: New item format specification 2. General Exam instructions 3. Sent/attached documents 4. Important Exam notifications 5. The appearance of Exam screen to the candidate	1. E.g., image enlargement, font size, colour and background, abbreviations. 2. A unique identifier, marked/unmarked questions, one-time exam, upload on STP system, upload data for all items, no examinee image to appear only ID Check, examinee image is taken on the test site. 3. Word doc. Excel, and what it contains 4. The appearance of question same as submitted, same sequence as submitted, 5. A) confirmation of the name of the examinee and exam with a question are you (name)? On the first page, B) appearance of screen to the examinee (question to appear on one screen, if too long (over the standard screen size) then option of scrolling and not splitting of the screen), video to be embedded within the question and not to appear as an exhibit. C) Button options: inclusion of buttons (skip, previous, next, review and mark). The functionality of a close button to go back to the exam e.g., from the review screen, basic calculator button to appear with relevant questions, Comment button functionality for candidates' comments A warning to appear in the middle of the screen if the exam was ended accidentally, or if there were unanswered, unmarked or skipped questions.

4.	Review Disk Information	Who reviews the disk, any specifications required	E.g., the disc should contain both exam forms sequentially as it should appear on the day of delivering the live exam.
5.	Exam Delivery	Details of exam site, date, time and client	International, July 8, 9:00 am, SCHFS
6.	Content delivery and Item banking	Mode of the provision of items, provision of unique identifier, number of files delivered.	e.g., items to be delivered electronically, unique master code, 2 files.
7.	Form specification	Number of forms delivered, name, the ID of form, overall time, status (e.g., sequential), marked and unmarked items	E.g., two forms, name: schs_7768848 Status: New form
8.	Screen section information	Information about the screens to be viewed by the examinee, setting if it is to be optional or required	e.g., schs_7768848 160/130 show timer, required, screen tutorial: required, show timer (optional).
9.	Additional information	Any extra requirements needed to be included in the exam screen	e.g., client logo, calculator, examinee image, the appearance of a digital clock, as well as a warning after an hour and when there are 10 minutes left of the exam.
10.	Functions of exam screen	Standard buttons and any additional buttons to include	e.g., mark, scroll down, next, skip, back, calculator (basic, scientific or custom), hotkeys enabled, others: e.g., abbreviations included at bottom of the question
11.	Functions for examinees	Additional functions to be available during the exam	Review items at any time during the exam, comments on items at the end of the exam only
12.	Exam presentation (style)	Font size, style, and colour, Background colour, margin details	Item format: Sans serif text (Calibri), font size between 20-24, word count per line $\leq$8. This does not count (prepositions, possessive pronouns, indefinite articles or numbers within the text).
13.	Item information	Item Type	MCQs – single response, video .avi files
		Pop-up exhibits	Some video may need to be viewed as an exhibit
		Graphics (gif, jpg, etc.), special characters	Two images to appear next to each other
		Language, Type	English UK the content and non-content section, tile bar, buttons. Text alignment to the left.

	Logo on examinee document	e.g., logo appearance on approval screen, and at the end.
	Screen resolution	Resolution: default 1024X768
	Result type Availability of cut score	Simple XML – Results, as well as image, post-administration determined
14. Score information	Percent score	% correct on total item
	Rounding	none
	weight of item	Dichotomous (0/1), no -ve mark
	Score report Category	Overall for the entire exam Number correct by category/domains
	Score report information	Notice of exam completion, not to appear to examinees but to be allowed for printing by the TCA No pass/fail to appear
	Examinee sore report (select if to be displayed, printed or mailed)	Score report for scored items only and of category and performance details as percent correct, percent correct by category, number correct by category and number of items delivered by category,
	Demographic information to be displayed on the screen, logo and others.	Examinee demographic information (name, speciality, government ID, passport, exam centre, region, eligibility ID)
15. Client information	Reporting information for the client	Repot description by section and category
	Result information for the client	Client results were received through STP, results were not submitted to a third party, results file included examinees' comments
16. Attachments and Add-ons	Any additional information that needs to be relayed through the form of an attachment	Name and type of document e.g., Item format specification and EM promotion BP in the form of ".doc", size 45-megabyte, View link to attachment

Appendix 18: Cambridge framework

Cambridge framework adapted from Shaw S, Crisp V, Johnson N

Inference and Warrant	Validation question	Possible methods
Construct Representation Tasks elicit performances that represent the intended constructs	1.Do the tasks elicit performances that reflect the intended constructs?	Review examiner reports for insights into how the questions were answered by candidates.
		Analyse performance data (e.g., item level scores) for a sample of candidates using statistical methods (e.g., Rasch, factor analysis) to explore item functioning, relationships between items, and to check for test bias (e.g., using differential item functioning analyses by gender, school type, etc.).
		For misfitting items, analyse the nature of candidate responses to gather insights into any possible sources of construct irrelevant variance.
		Ask appropriate examiners/experts to rate the extent to which each question appears to elicit each assessment objective set out in the syllabus (using this as a proxy for the constructs).
		Ask appropriate examiners/experts to rate the extent to which each question places certain types of cognitive demands on students.
Scoring Scores reflect the quality of performances on the assessment tasks	2. Are the scores/grades dependable measures of the intended constructs?	Review exam board documents on marking and scoring procedures.
		Ask a number of markers to mark the same exam scripts in a multiple re-marking exercise so that the consistency and reliability of marking can be analysed.
		Conduct statistical analyses of candidate exam results to explore issues relating to aggregation of test scores and intended and achieved weightings of exam components. •
		Conduct composite reliability analysis. •

		Statistical analysis of the effectiveness and accuracy of classifying students to grade bands based on marks.
Generalization Scores reflect likely performance on all possible relevant task	3.Do the tasks adequately sample the constructs that are set out as important within the syllabus?	Ask appropriate examiners/experts to identify the topics and sub-topics assessed by each exam question for a number of exam sessions in order to evaluate content and skills coverage.
		Ask appropriate examiners/experts to rate for each exam question the cognitive demands rewarded by each question, as reflected by the mark scheme.
		Ask appropriate examiners/experts to rate for each exam question the extent to which the scoring guidelines set out in the mark scheme reward each assessment objective.
Extrapolation Score reflect likely wider performance in the domain	4.Are the constructs sampled representative of competence in the wider subject domain?	Ask higher education representatives and employer representatives to review the syllabus content in relation to the preparation it provides for further study or employment.
		Conduct longitudinal studies involving correlations between test scores/grades and performance in subsequent education or employment.
Decision- making Appropriate uses of scores are clear	5. Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?	Review available guidance documents.
		Use a questionnaire to teachers to gather their views on guidance on score/grade meaning and uses and gather insights on how they use scores/grades.
		Use a questionnaire to stakeholders (e.g., higher education providers) to gather their views on guidance on score/grade meaning and uses and gather insights on how they use scores/grades.

Shaw S, Crisp V, Johnson N. A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assessment in Education: Principles, Policy & Practice. 2012;19(2):159-76. (35)

Appendix 19: Checklist for evaluating a mixed-methods research study

Appendix 19 with the table of **Checklist for evaluating a mixed-methods research study** has been removed due to Copyright restrictions.

Adopted from Johnson RB, Christensen L. Educational Research: Quantitative, Qualitative, and Mixed Approaches. 2016. (159 p.104),

Appendix 20: Validity checks for quantitative, qualitative and mixed methods research

Validity check for quantitative research methods (159) p284-92

Type	Meaning	Example	Comment
Internal Validity (Causal Validity)	Establishing evidence that a relationship is present between two variables (IV and DV)	**Maturation**: (a mental or physical change) over time that may have an effect on performance on the DV (e.g., learning, boredom, hunger, fatigue)	This can be overcome by having a control group because the effect of the threat occurs for both groups equally.
		Differential selection: differences in characteristics between participants (age, gender, ability, intelligence, language, learning style, maturity, motivation, anxiety, stress, etc.)	Overcome the threat of differential selection through the use of random assignment that equates the groups (in a large enough sample size) and therefore, any differences that occur are due to the independent variable.
External validity (Generalizing Validity)	The extent that the study results can be generalized to other populations, times, and setting	**Population validity:** generalizing the study from sample to population and to other types of people in the target population.	G-Study was conducted, although G-coefficient was not high enough, it could be due to the low number of items. Further reliability studies would be needed.
		Ecological validity: generalizing results across any settings **Reactivity effect:** changes in performance due to participants knowing that they are part of a study	This is not applicable based on the results of this study. Further similar studies need to be conducted fist A few residents commented that they felt uncomfortable not knowing which item was marked. However, this could not be assessed.
		Temporal validity: generalizing results across time	Would need further reliability studies
Construct Validity	The extent that the higher-order construct is accurately Measured and "operationalized" in a study	**Treatment diffusion:** participants interact or share resources with the other group	Identified the construct (higher cognitive skills) and provided an explanation
Statistical Validity	The extent to which the relationships between the variables (independent and dependent) are related in the larger population	**Effect size:** measures the strength (magnitude) of the relationship between the variables	Statistical analysis of the items and their relationships were provided. No correlation studies were taken between results and other methods measuring the same construct.

Validity check for qualitative research methods (159) p284-92

Type	Meaning	Example	Comments
Trustworthiness	Validity in qualitative research that is plausible, credible and defensible	E.g., presenting all views, strength and weaknesses of methods, data collection, residents' opinions	Providing details and evidence throughout the research, as well as challenges and limitations.
Triangulation	A validation approach using multiple methods, investigators, data sources when searching for results convergence.	Using test, focus group and questionnaire, gathering multiple perspectives from different regions to provide a better perspective and understanding of the phenomenon.	Multiple methods were used (test, pilot, focus group, questionnaire) and multiple data sources (collecting data from multiple sources, at different times and places)
Research Bias	Collecting results that the research wants to find	Allowing one's personal views affect in Selecting recording and interpreting the information	Overcome by reflexivity: the researcher actively participates in self-reflection about his/her perspective and biases. And negative-case sampling: searching for examples that disconfirm the researcher's expectation about the study.
Descriptive Validity	The accuracy in reporting descriptive information (facts, events, setting, behaviours, etc.) by the researcher	Is it accurate? what is reported: - What actually happened - What was seen and heard	Detail reports, thematic analysis and narrative review throughout the research
Interpretive or Emic Validity	Accurately reporting and presenting participants' meaning, viewpoints, perspectives, thoughts, feelings and their subjective worlds	Using feedback to clarify misconceptions and understand their feelings.	Participant feedback was presented, checking with residents to clear up miscommunications and inaccuracies about what they said In transcription and analysis: used verbatim descriptions that are phrased very similarly to what residents meant to relay it to the reader. Provided direct quotations pf participants' exact words, as well as their actions and feelings.
Theoretical Validity	The degree to which the developed theoretical explanation of the phenomena fits the data making it defensible and credible	Explaining the phenomena and why it operates as it does. The theoretical construct of the thinking process of scenarios is used to explain the students' perception to	Strategy to further support this is to do extended fieldwork and collecting data from the field over a period.

		the MM-TXT items, diagnosis, and choices of answer selection. (other factors play a role: break, fatigue, CBT, learning preferences).	The use of multiple theoretical perspectives, disciplines to interpret and explain the data. Cognitive load theory, multimedia learning, test-taking strategies. Results match to what was predicted that multimedia items are more discriminating and sometimes more difficult. Peer review of results through discussion with colleagues and the involvement of a 'Critical friend' who interacted with the research throughout the whole process for feedback regarding actions
Internal validity	The degree to which the researcher describes the phenomena of how it operates and obtains a causal explanation and tests its theory	Using critical friend and multiple methods, triangulation and reflexivity.	To improve internal validity in a qualitative study, one can use the following strategies: Critical friend, low-inference descriptors, multiple data source, multiple methods, multiple theoretical perspectives, negative-case sampling (where results didn't match), participant feedback, member checking, reflexivity, Triangulation, peer review, ruling out alternative explanations. All these were used.
External validity	Generalizing the findings to other people, places and times.	In qualitative research, this is considered a weakness because people are not randomly selected and it usually focuses on documenting particular findings in a certain context rather than a universal one.	Naturalistic Generalization could be evaluated with other specialities: The more similar people are in characteristics and circumstances the more defensible the generalization can be Provide all information about the participants in the study, the selection, context, setting, methods, data collection and analysis to give the reader the ability to make the decision of generalizability and to replicate the study with new participants (replication logic).

Validity check for mixed-methods research based on Onwuegbuzie and Johnson's 2006 types of mixed research validity (also called types of legitimation)(226 p.306-9)

Type	Meaning	Example	Comment
1) inside-outside legitimation	The extent to which the researcher accurately reports understands and presents participants' meaning, viewpoints, perspectives, thoughts, feelings and their subjective worlds (emic viewpoint), as well as the researcher's outsider view (etic viewpoint)	Presented quotes in transcriptions and feelings towards items and CBT	Fully present both perspectives the participants (through transcription) and the researcher (reflexivity) and move back and forth between these viewpoints.
2) paradigmatic/philosophical legitimation	The degree to which the researcher reflects on, understand and clearly explains the philosophical beliefs about the mixed-method research undertaken	Demonstrate an understanding of why the quantitative and qualitative aspects were undertaken from a philosophical view.	Fully explain the philosophical and methodological paradigm including the researcher's epistemological, ontological, methodological, and beliefs about mixed research
3) Commensurability approximation legitimation	The degree to which the researcher can take the quantitative and qualitative views and integrate them into a mixed broader viewpoint	Look through combined or multiple-lenses to have a better understanding of explaining the phenomenon	Strategies used are: The researcher must think, become and look through the lens as a quantitative researcher and as a qualitative researcher and move back and forth thill he/she becomes a mixed researcher (Gestalt switches) Role reversal and empathy towards examinees

4) Weakness minimization legitimation	The degree to which one research approach's weakness is compensated by the strength of another research approach.	IA doesn't reflect the whole reason and is compensated with focus group feedback. G-theory replaces CTT reliability deficit.	Combine qualitative and quantitative approaches that have non-overlapping weaknesses. Using test, pilot, focus group and questionnaires and triangulation of results from all.
5)Sequential legitimation	The degree to which the researcher appropriately addresses and builds on one phase of the research design on the other and (from the qualitative and quantitative phases)	Building the FG based on the results of the MM-TXT matched item analysis and the questionnaire results.	The researcher needs to understand which research design needs to be conducted and purposively builds the second phase based on the findings from the earlier phases, thus achieving sequential validity. Here a QUAN-QUAL sequential design was taken.
6) Conversion legitimation	The degree to which the researcher makes high-quality data transformation and based on it makes appropriate interpretation	Quantitizing (quantifying qualitative data) or Qualitizing (reporting quantitative data into words, themes and categories)	The researcher provided quality inferences from quantitizing and qualitizing some of the data from IA and questionnaires, as well as presenting quantitative analysis for focus group discussions.
7) Sample integration legitimation	The degree to which the relationship between the QUAN and QUAL sampling design, the conclusion, generalization and meta-inferences are appropriately made by the researcher	IA reflected that some MM items took longer hence may be more difficult. However, in the focus group discussion, residents attributed longer duration in some items to quality of media and repeating videos even if it was easy.	The researcher should keep in mind that the participants in the QUAN and QUAL groups may not have the same beliefs and has to be careful in combining the data Strategies to use is to study the degree to which the research purpose was met, the problem under study was solved and the research question was answered. (i.e., pragmatic legitimation) Integration legitimation: the degree to which the QUAN and QUAL data were integrated, analysed

			and concluded have been achieved by the researcher.
8) Socio-political legitimation	The degree to which the researcher appropriately addresses the multiple viewpoints of the involved stakeholders	Certain rules and regulation adopted in the organisation may restrict the optimal method of test administration (e.g., method of standard-setting, using only CTT and not IRT), no set break available.	The researcher fully understood and represent interests, values and viewpoints related to the research topic
9)Multiple validities legitimation	The degree to which the researcher addresses and successfully resolves all the relevant validity types mentioned in the QUAN, QUAL and mixed methods research	Reviewing all types of QUAN and QUAL validity mentioned in previous tables.	The most important type of validity in mixed research. The researcher must identify what are the relevant types of validity, trustworthiness and legitimations that need to be addressed in the study.
Pragmatic legitimation	The extent to which the research purpose was met, problem solved, questions were answered and results provided.	Problem: first-time takers for CBT, solution: orientation, presentation and material distributed. Problem: non-random sampling of examinees, solution: random assignment	The researcher presented results to questions, explained weaknesses and strength of validity framework and methods to overcome problems (in results chapter 6) To what extent does the research motive others to use the findings.
Integration legitimation	The degree to which QUAN and QUAL results are combined in a third viewpoint	e.g., interpretation form IA to its reason from the questionnaire and focus group	QUAN and QUAL results were presented in the results chapter separately and integrated as a whole in the discussion.
Research reliability	Same results are obtained if the research was repeated	If research repeated on other EM residents with other items would the results be the same?	Reliability study was undertaken, and the results were good. Although not strong probably due to not having a larger sample size. Requires further replication of study and observing the consistency of results.

Appendix 21: Example of a combined result of QUAN and QUAL analysis

Example of a combined result of QUAN and QUAL analysis

A healthy 56 year-old woman presented with left facial, arm and leg pain of one hour's duration (see scan).

Blood Pressure	195/115 mmHg
Hear Rate	95/min
Respiratory Rate	23/min
Temperature	36.0° C
Oxygen Saturation	96% on face mask

What is the best next step in management?

A. t-PA
B. Labetolol
C. Lumbar puncture
D. Admission for monitoring

Item Parameters

DIFF	DI	RPB	KEY
0.55	0.15	0.38	B
A	B	C	D
11.4	54.5	6.8	18.2

Item Description

A question with moderate difficulty that does not correlate with students' overall performance but still can discriminate between them.

A healthy 56 year-old woman presented with left facial, arm and leg pain of one hour's duration.

Blood Pressure	195/115 mmHg
Hear Rate	95/min
Respiratory Rate	23/min
Temperature	36.0° C
Oxygen Saturation	96% on face mask

Scan: right hyperdense middle cerebral artery sign

What is the best next step in management?

A. t-PA
B. Labetolol
C. Lumbar puncture
D. Admission for monitoring

Item Parameters

DIFF	DI	RPB	KEY
0.73	0.32	0.52	B
A	B	C	D
13.6	72.7	2.3	2.3

Item Description

A very good question with moderate difficulty and discriminates between students and correlates with their overall performance.

Explanation by Focus group discussion: why MM item was less discriminating in this item and more difficult	**Theme**
R1-G: *no it's not that clear by the way,* I would miss it R1-N: *on the contrary, for me, this* image helped me more, *to be honest* R1-A: *I think without the picture, the answer will be clear.*	**Clarity of MM**
R1-N: *maybe on the screen, it's more clear*, *it's more clear on the screen*	**CBT - screen**
R1-G: *No this is unfair for* R1 V-BH: *no if he brought a* clear CT, *the* junior he has to know, *this one* V-5: *and this* promotion *will define the* R1 *and* R2 *and* R3	**Difficulty level- level or residents**
R1-G: *no* description *is better*	**Item Format – Preference**

V-G1: *but* still, *the* text *compared to the image here, the* text *is easier* V-NP: *yes, it was confusing a lot this is what confused me; without the* CT *it was very* easy	**Presence/absence of MM**
R1-G: *so, for me, the* description *became easier than the* CT, *the time that I will sit reading the* CT *it takes time*, *it's not like when I read a* description, *it's just faster*.	**Time – Importance of time**
R1-V: *the idea is, the* image first you will see, interpret the image, after that you will see there is early sign of ischemic stroke, then you will think I should give tPA but there is contraindication by the vital, the blood pressure, so I should decrease the blood pressure before that. Then I will give the tPA, so first choice will be to give labetalol which will reduce the blood pressure. While in the other question in the written one, you will see the right hyperdense middle cerebral artery its infarction, you will go to the vital directly, it's high, give labetalol directly. R1-N: *on the contrary, for me this* image helped me more *to be honest*, *because this* hyperdense *is going to make me think of* bleeding, *it's right that in the end, the answer is going to be the same but the questions this time, is the same answer but maybe there is another question…* I'm not focusing per level, *when I first saw the question, I'm an* R3 *and I'm talking now, when I first saw the question I knew that he wanted* labetalol, even *before I reached it*, *as soon as I read the* blood pressure 195, *I knew he needed something to decrease it, but when I read* hyperdense *I said why not* bleeding. *Then I said if it's* bleeding, *they're not going to put a high blood pressure then come back and tell me* bleeding. *The questions "*next step" *this is what cleared it more for me that he wanted to decrease the pressure then give* tPA. Ok *but if someone didn't pick this up or he knows this is* common, *the consultants always asked us in it and* hyperdense *will come with* bleeding. *What is the most thing that is* hyperdense *in a* CT? R1-G: bleed R1-N: bleed, *so* bleeding *anyways will change all the answers*	**Direction of thoughts**
V-2: *exactly*, *the* CT *will not benefit me with anything because* it will not change the management V-BL: *the* history is, history is not correlated *with the* image	**Relevancy of MM**
V-K: *but the problem even it didn't say* weakness *it said* pain *so it makes you confused a bit if it was* weakness *it would be* ok *you look for* signs V-S: *but there is another thing in the Q that is* confusing, *that it is* left facial arm and leg pain. It presents with weakness *more than* pain *so this was…*	**Language – choice of words**
V-F: it's a hyperdense *it's* acute bleeding V-J: hyperdense *means* bleed V-W: a specific sign, *it has a certain* terminology, hyperdense MCA *sign*	**Cue and Clues**
V-W: *but the* image, *you won't know it if you never saw it*	**Exposure**
V-G1: *that* bleeding, something big, *but here I feel that I don't know*	**Severity of the condition, Clarity of MM**
R1-Ali: *if this image, for instance, had an* area, they mark the area, *maybe it would be* more fair *for* juniors	**Labelling** **Difficult – level of resident**