VALIDATION OF SIMULATION AS AN ASSESSMENT TOOL FOR ASSESSING COMPETENCE OF MEDICAL STUDENTS AND JUNIOR DOCTORS IN MANAGING CARDIAC EMERGENCIES

Dr Anastasia Vlachadis Castles
BMBS, BRadMedImag (Hons), DipLang

(Student Number: 21544591)

This thesis is presented in partial fulfilment of the requirements for the Master of Health Professional Education (Thesis and Coursework) of the University of Western Australia.

Graduate Research School,
Faculty of Medicine, Dentistry and Health Sciences
University of Western Australia

2016
ABSTRACT

Background: Junior medical officers are expected to be able to manage and lead cardiac emergency situations with confidence and efficiency, yet studies have demonstrated that medical graduates are underprepared to manage medical emergency situations. It is imperative that we are able to assess the competence of medical students and junior doctors in managing cardiac emergencies, to ensure that junior doctors are prepared for practice and are able to provide safe and efficient patient care in these situations.

This thesis contributes to our understanding of assessment in medical education, in particular the use of simulation as a method for practical assessment. The primary objective of this study was to validate and compare simulation-based assessments with multiple choice question tests (MCQ) as methods for assessing competence in cardiac emergency management, in a population of medical students and junior medical officers.

Methods: In order to answer the research questions in adequate detail, this study used a mixed methods approach. A prospective study design was used, with medical students and junior doctors recruited to complete an MCQ test and one of two alternate simulation-based assessments (live actor in a hybrid set-up). The quantitative data was analysed using descriptive and inferential statistical measures to examine the psychometric properties of the simulation assessments and MCQ tests, specifically, construct validity, formative validity and inter-observer reliability. Qualitative analysis was used to further explore formative validity and evaluate content validity, by descriptively analysing the assessment tools, MCQ feedback document, simulation
debrief instructions, and the field notes. As the field notes were acquired and reviewed, an abundance of data on the research process and simulation assessment was uncovered, in response to which examination of the simulation assessment research process was identified as an additional objective of the study. For this main qualitative component, a grounded theory approach was used: the field notes were analysed using inductive coding, with the codes and subcodes identified then used to develop a taxonomy describing the domains and dimensions involved in the simulation assessment research process.

**Results:** Thirty-eight participants (17 students and 21 junior medical officers) were recruited. All participants completed the MCQ test, and 37 participants completed one of two alternate simulation scenarios. Results for the MCQ test and the overall simulation scores were normally distributed. A moderate and statistically significant correlation between results of the MCQ test and simulation assessments was found (r = 0.52 for Scenario 1, r = 0.57 for Scenario 2).

Qualitative analysis demonstrated that both assessment methods (simulation and MCQ) had adequate content validity, however learning objective categories were covered more evenly by the simulation assessments. In terms of construct validity, none of the assessment tools had acceptable internal consistency, with the MCQ test, Scenario 1 and Scenario 2 having coefficient alpha values of 0.64, 0.52 and -0.03, respectively. The MCQ test and Simulation Scenario 2 (but not Scenario 1) demonstrated adequate construct validity using the contrasted-groups approach, with statistically significant differences in performance between ‘novice’ and ‘expert’ groups on these two assessments.
Formative validity was adequate for the MCQ test, though this was limited by the inherent properties of this type of assessment. The simulation assessment tools, by virtue of their format and inclusion of a debriefing component, had very high formative validity. Interobserver reliability was very high for Simulation Scenario 1 and fair for Simulation Scenario 2, but was dependent on the combination of assessors.

**Conclusions:** Overall, the simulation assessment tools designed for this research project were shown to have similar psychometric properties to the MCQ test, which was used as a comparison. Each assessment method had strengths and weaknesses, but the most notable difference was the superior formative validity of the simulation assessments. Given that there was moderate correlation between the two assessment methods (simulation and MCQ), consideration could be given to employing simulation assessment in addition to the traditional MCQ assessment. Simulation would be best used as a formative assessment, so that examinee performance can guide further teaching and learning.
ACKNOWLEDGMENTS

Supervision
Supervisor:
Professor Sandra Carr
RN, RM, BSc (Curtin), MPH (W.Aust), PhD
Associate Dean (Education)
Director of Postgraduate Courses in Health Professional Education
Faculty of Medicine, Dentistry and Health Sciences, The University of Western Australia

Co-Supervisor:
Associate Professor William van Gaal
MBBS, MD (Melb), MSc (Oxon), FRACP, FCSANZ, FESC
Divisional Director of Emergency Services and Director of Cardiology
The Northern Hospital, Victoria
University of Melbourne

Assistance
Participant Recruitment:
Dr. Leonie Griffiths
Dr. Carol Chong
Dr. Louisa Lee
Amanda Geddes

MCQ Review:
Associate Professor William van Gaal
Professor Sandra Carr
Associate Professor Zarrin Siddiqui
Dr. Harendra Wijesekera
Dr. Leonie Griffiths
Dr. Carol Chong
Dr. Roza Nastovska
Dr. Josh Chatelier

Simulation Scenario Review:
Associate Professor William van Gaal
Professor Sandra Carr
Dr. It Meng (Tim) Tsay
Dr. Leonie Griffiths
Dr. Roza Nastovska
Dr. Josh Chatelier

Simulation Scenario Execution:
Dr. Simon Ho
Angelise Vlachadis Castles
Tones Pask
Nicole Garrow
Alicia McMahon
Simulation Scenario Assessment:
Dr. It Meng (Tim) Tsay
Dr. Henry Wong
Dr. David Ta
Dr. Simon Ho

Statistics Guidance and Review:
Centre for Applied Statistics, University of Western Australia

Thank you to all of those who offered support and assistance with this research project.

To Professor Carr: Thank you for guiding me through the process, for challenging me, and for reassuring me that the stressful moments were normal.

To Associate Professor van Gaal: Thank you for your time, advice and wisdom. Thank you for your ongoing mentorship, and for believing in me.

To those who helped with reviewing my work, recruiting participants, and running/assessing the simulations: Thank you for being so giving of your time. Thank you for all your effort, and for helping to make this project a success. I could not have done it without you.

To my friends: Thank you for encouraging me when it all felt like too much. Thank you for being my relief from study. Thank you for staying around when I had too many priorities to juggle.

To my family: Mum, Dad, Angelise- You have given me the strength and belief that drives me in everything I do. You have made me feel like I can achieve anything, and then always been there to help me along the way. Thank you for your unconditional love and support- it means the world to me!
STATEMENT OF CANDIDATE CONTRIBUTION

This thesis does not contain work that I have published, nor work under review for publication.
TABLE OF CONTENTS

ABSTRACT ......................................................................................................................... i
ACKNOWLEDGMENTS ....................................................................................................... iv
STATEMENT OF CANDIDATE CONTRIBUTION .......................................................... vi
ABBREVIATIONS AND ACRONYMS ........................................................................... 1
LIST OF TABLES ............................................................................................................... 2
LIST OF FIGURES ............................................................................................................. 4

SECTION I: INTRODUCTION .......................................................................................... 5
CHAPTER 1: BACKGROUND .............................................................................................. 5
A. Introduction .................................................................................................................... 5
B. Requirements of Junior Doctors .................................................................................. 7
C. Choice of Assessment Methods .................................................................................. 10
D. Simulation .................................................................................................................... 16
E. Differentiating Simulation Assessment from Objective Structured Clinical Examinations .................................................................................................................................................. 17
F. Evaluation of Assessment Tools .................................................................................. 19
G. Objectives ..................................................................................................................... 25
H. Research Questions ..................................................................................................... 26
I. Summary and Thesis Structure ................................................................................... 27

CHAPTER 2: LITERATURE REVIEW ................................................................................ 29
A. Introduction .................................................................................................................... 29
B. Medical Graduates’ Preparedness to Manage Emergency Situations ......................... 30
C. Advantages of, and Barriers to, Use of Simulation as a Teaching Method in Medical Education .................................................................................................................................................. 31
D. Comparison of Simulation with Other Teaching Methods ........................................ 35
E. Assessing Performance of Examinees on Simulations ............................................... 36
F. Benefits of Simulation as an Assessment Method ....................................................... 37
G. Simulation in Teaching and Assessing Cardiac Emergency Management ................ 38
H. Studies Evaluating Psychometric Properties of Simulation Assessment Tools .......... 39
I. Studies Describing Simulation Assessment Tool Development .................................. 47
J. Conclusions ................................................................................................................... 49
K. Summary ...................................................................................................................... 51

SECTION II: METHODOLOGY ....................................................................................... 52
CHAPTER 3: METHODS OVERVIEW ............................................................................. 52
A. Research Approach ...................................................................................................... 52
B. Research Process ........................................................................................................ 55
C. Research Ethics ........................................................................................................... 66
D. Summary ...................................................................................................................... 68

CHAPTER 4: ASSESSMENT TOOL DEVELOPMENT ...................................................... 69
A. Multiple Choice Question Test Development .............................................................. 69
B. Simulation Assessment Development ......................................................................... 73
C. Summary ...................................................................................................................... 78

CHAPTER 5: ANALYSIS OF PARTICIPANT CHARACTERISTICS AND ASSESSMENT RESULTS ............................................................................................................. 79

CHAPTER 6: CONTENT VALIDITY ................................................................................. 82

CHAPTER 7: CONSTRUCT VALIDITY ............................................................................ 85
A. Internal Consistency .................................................................................................... 85
B. Contrasted-Groups Approach ..................................................................................... 86
C. Comparison Across All Groups ................................................................................ 87
CHAPTER 8: FORMATIVE VALIDITY ................................................................. 89
  A. Methods for Evaluating Formative Validity ............................................. 89
  B. Quantitative Component ......................................................................... 90
  C. Qualitative Component ......................................................................... 91
CHAPTER 9: INTEROBSERVER RELIABILITY ............................................... 92
CHAPTER 10: QUALITATIVE ANALYSIS OF THE RESEARCH PROCESS ......... 94

SECTION III: RESULTS .............................................................................. 96
CHAPTER 11: PARTICIPANT CHARACTERISTICS AND ASSESSMENT RESULTS .................................................................................. 96
  A. Participant Characteristics ...................................................................... 96
  B. Multiple Choice Question Test ................................................................. 98
  C. Simulation Assessments .......................................................................... 99
  D. Correlation Between MCQ Test and Simulation Assessments .................. 102
CHAPTER 12: CONTENT VALIDITY ............................................................... 104
  A. Multiple Choice Question Test ................................................................. 104
  B. Simulation Assessments .......................................................................... 107
CHAPTER 13: CONSTRUCT VALIDITY ......................................................... 111
  A. Multiple Choice Question Test ................................................................. 111
  B. Simulation Assessments .......................................................................... 116
CHAPTER 14: FORMATIVE VALIDITY .......................................................... 125
  A. Multiple Choice Question Test ................................................................. 125
  B. Simulation Assessments .......................................................................... 130
CHAPTER 15: INTEROBSERVER RELIABILITY ............................................. 144
CHAPTER 16: QUALITATIVE ANALYSIS OF THE RESEARCH PROCESS ...... 148

SECTION IV: DISCUSSION ........................................................................ 154
CHAPTER 17: PARTICIPANT CHARACTERISTICS AND ASSESSMENT RESULTS .................................................................................. 154
  A. Participant Characteristics ...................................................................... 154
  B. Multiple Choice Question Test ................................................................. 156
  C. Simulation Assessments .......................................................................... 157
  D. Correlation Between MCQ test and Simulation Assessments .................. 161
CHAPTER 18: CONTENT VALIDITY ............................................................... 163
  A. Multiple Choice Question Test ................................................................. 163
  B. Simulation Assessments .......................................................................... 168
  C. Comparison of the Multiple Choice Question Test and Simulation Assessments ............................................................... 169
CHAPTER 19: CONSTRUCT VALIDITY ......................................................... 173
  A. Multiple Choice Question Test ................................................................. 173
  B. Simulation Assessments .......................................................................... 183
  C. Summary of Construct Validity ............................................................... 189
  D. Comparison of the Multiple Choice Question Test and Simulation Assessments ............................................................... 190
CHAPTER 20: FORMATIVE VALIDITY .......................................................... 192
  A. Multiple Choice Question Test ................................................................. 192
  B. Simulation Assessments .......................................................................... 208
  C. Comparison of the Multiple Choice Question Test and Simulation Assessments ............................................................... 221
CHAPTER 21: INTEROBSERVER RELIABILITY ............................................. 224
CHAPTER 22: QUALITATIVE ANALYSIS OF THE RESEARCH PROCESS ...... 233
  A. Participant Recruitment and Logistics ....................................................... 233
  B. Simulation Assessment Execution ............................................................ 238

SECTION V: SUMMARY AND CONCLUSION .............................................. 267
CHAPTER 23: SUMMARY OF THE RESEARCH FINDINGS .............................. 267
CHAPTER 24: RECOMMENDATIONS .......................................................... 272
CHAPTER 25: LIMITATIONS ....................................................................... 278
CHAPTER 26: FUTURE DIRECTIONS ............................................................ 285
CHAPTER 27: CONCLUSION ....................................................................... 289
REFERENCES ............................................................................................. 292
Appendix 1: Participant Information Form and Participant Consent Form .......... 1
Appendix 2: Participant Questionnaire ......................................................... 8
Appendix 3: Multiple Choice Question Test ................................................... 11
Appendix 4: Multiple Choice Question Test Feedback ................................. 16
Appendix 5: Simulation Scenarios ................................................................. 25
Appendix 6: Recruitment Material ................................................................. 55
# Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>Australian Curriculum Framework</td>
</tr>
<tr>
<td>AMI</td>
<td>Acute myocardial infarction</td>
</tr>
<tr>
<td>APO</td>
<td>Acute pulmonary oedema</td>
</tr>
<tr>
<td>BMBS</td>
<td>Bachelor of Medicine/Bachelor of Surgery course</td>
</tr>
<tr>
<td>CK</td>
<td>Creatine kinase</td>
</tr>
<tr>
<td>CXR</td>
<td>Chest X-ray</td>
</tr>
<tr>
<td>Dx</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>Ex</td>
<td>Examination</td>
</tr>
<tr>
<td>FBE</td>
<td>Full blood examination</td>
</tr>
<tr>
<td>HMO2</td>
<td>Hospital medical officer- year two</td>
</tr>
<tr>
<td>Hx</td>
<td>History</td>
</tr>
<tr>
<td>IHD</td>
<td>Ischaemic heart disease</td>
</tr>
<tr>
<td>Ix</td>
<td>Investigation(s)</td>
</tr>
<tr>
<td>JMO</td>
<td>Junior medical officer</td>
</tr>
<tr>
<td>MBBS</td>
<td>Bachelor of Medicine/Bachelor of Surgery course</td>
</tr>
<tr>
<td>MCQ</td>
<td>Multiple choice question</td>
</tr>
<tr>
<td>MD</td>
<td>MD (Doctor of Medicine) course</td>
</tr>
<tr>
<td>MD2</td>
<td>Second year of the University of Melbourne MD program</td>
</tr>
<tr>
<td>MD3</td>
<td>Third year of the University of Melbourne MD program</td>
</tr>
<tr>
<td>MD4</td>
<td>Fourth year of the University of Melbourne MD program</td>
</tr>
<tr>
<td>OSCE</td>
<td>Objective structured clinical examination</td>
</tr>
<tr>
<td>PGY1</td>
<td>Postgraduate year one (intern)</td>
</tr>
<tr>
<td>PGY2</td>
<td>Postgraduate year two</td>
</tr>
<tr>
<td>UEC</td>
<td>Urea, electrolytes, creatinine</td>
</tr>
<tr>
<td>UoM</td>
<td>University of Melbourne</td>
</tr>
<tr>
<td>UWA</td>
<td>University of Western Australia</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1.1: Krathwohl’s Taxonomy Table (2002)(18) .......................................................... 12
Table 2.1: Most commonly evaluated psychometric properties and methods used to measure these properties .......................................................... 45
Table 2.2: Research methods and major findings of the studies evaluating psychometric properties of simulation assessment tools ........................................ 46
Table 11.1: Demographic data of participants overall, and of the groups allocated to Scenario 1 and Scenario 2 ............................................................. 97
Table 11.2: Information on participant medical training ..................................................... 97
Table 11.3: Descriptive statistics for the simulation assessments ...................................... 99
Table 11.4: Correlation between the checklist and GAS components of the simulation assessment tools ........................................................................ 101
Table 11.5: Correlation between the MCQ test and simulation assessments, including each component of each simulation assessment ...................................... 102
Table 12.1: Learning outcomes for AMI and APO, and corresponding MCQ items .......................... 105
Table 12.2: Keywords identified for each MCQ item ................................................................ 106
Table 12.3: Learning outcome categories for AMI and corresponding simulation assessment items (Scenario 1) ............................................................... 107
Table 12.4: Mapping of simulation assessment items (Scenario 1) to specific learning outcomes ......................................................................................... 108
Table 12.5: Learning outcome categories for APO and corresponding simulation assessment items (Scenario 2) ................................................................. 109
Table 12.6: Mapping of simulation assessment items (Scenario 2) to specific learning outcomes ......................................................................................... 110
Table 13.1: Internal consistency statistics for the MCQ test ............................................... 111
Table 13.2: Cronbach’s alpha for each MCQ item (all items considered as a single assessment tool) ........................................................................... 112
Table 13.3: Cronbach’s alpha for each item on the MCQ test (items divided by topic and considered as two separate assessment tools) ......................... 113
Table 13.4: Kruskal-Wallis comparison pairwise across all classifications for MCQ scores ................................................................................................. 116
Table 13.5: Internal consistency statistics for the simulation assessments ......................... 116
Table 13.6: Cronbach’s alpha for each item on Simulation 1 ................................................ 117
Table 13.7: Cronbach’s alpha for each item on Simulation Scenario 2 ............................ 118
Table 13.8: Comparison of simulation component scores for the ‘novice’ and ‘expert’ groups ............................................................................................... 121
Table 13.9: Comparison of simulation component scores across all classifications .......... 124
Table 14.1: Mean item value and standard deviation for each item of the MCQ test ........ 125
Table 14.2: Spread of item difficulty values for the MCQ test ............................................. 126
14.3: Topics and categories of learning outcomes covered by MCQ items, sorted by difficulty ......................................................................................... 127
Table 14.4: Results for MD2 and HMO2 participants for MCQ items 8 and 9 ................. 128
Table 14.5: Results for participants of all classifications for MCQ item 25 ................. 128
Table 14.6: Kruskal-Wallis comparison pairwise across all classifications for MCQ Item 25 ................................................................................................. 128
Table 14.7: Mean item values and standard deviation for each item of Simulation Scenario 1 ........................................................................................................ 131
Table 14.8: Spread of item mean item values for Simulation Scenario 1 ......................... 132
Table 14.9: Topics and categories of learning outcomes covered for Simulation Scenario 1, sorted by mean item value ......................................................... 133
Table 14.10: Results for MD3 and HMO2 participants for Simulation Scenario 1 items A1c and A4a ................................................................. 134
Table 14.11: Results for participants of all classifications for Simulation Scenario 1 items A1c and A5d ................................................................. 134
Table 14.12: Mean item value and standard deviation for each item of Simulation Scenario 2 .................................................................................. 136
Table 14.13: Spread of mean item values for Simulation Scenario 2 .................. 137
Table 14.14: Topics and categories of learning outcomes covered for Simulation Scenario 2, sorted by mean item value .................................................... 138
Table 14.15: Results for MD2 and HMO2 participants for Simulation Scenario 2 items A4a and A4c ........................................................................ 139
Table 14.16: Results for participants of all classifications for Simulation Scenario 2 items A4a, A4b, A4c and A4d ................................................................. 140
Table 15.1: Percentage agreement for each combination of assessors ............... 144
Table 15.2: Spearman correlation for each combination of assessors .................. 145
Table 15.3: Cohen’s kappa for each combination of assessors ............................ 146
Table 15.4: Percentage agreement for each item .................................................. 147
LIST OF FIGURES

Figure 1.1: Miller's Pyramid of Assessment ............................................................... 13
Figure 3.1: Research process outline .......................................................................... 55
Figure 3.2: Allocation of participants to simulation scenarios .................................... 60
Figure 11.1: Distribution of scores for the MCQ test ................................................. 98
Figure 11.2: Distribution of total scores for Simulation Scenario 1 .......................... 100
Figure 11.3: Distribution of total scores for Simulation Scenario 2 .......................... 100
Figure 11.4: Correlation between checklist score and GAS for Scenario 1 ............. 101
Figure 11.5: Correlation between checklist score and GAS for Scenario 2 ............. 102
Figure 11.6: Correlation between the MCQ test and simulation assessments (Scenario 1) .................................................................................................................................................. 103
Figure 11.7: Correlation between the MCQ test and simulation assessments (Scenario 2) .................................................................................................................................................. 103
Figure 13.1: Boxplot comparing MCQ scores for MD2 and HMO2 groups .......... 114
Figure 13.2: Boxplot comparing MCQ scores across all classifications .................. 115
Figure 13.3: Boxplot comparing simulation assessment scores for Scenario 1 for MD3 and HMO2 groups ............................................................................................................................................ 119
Figure 13.4: Boxplot comparing simulation assessment scores for Scenario 2 for MD2 and HMO2 groups ............................................................................................................................................ 120
Figure 13.5: Boxplot comparing simulation assessment scores across all classifications for Simulation Scenario 1 (total score) ........................................................................................................................................... 122
Figure 13.6: Boxplot comparing simulation assessment scores across all classifications for Simulation Scenario 2 (total score) ........................................................................................................................................... 123
Figure 14.1: Spread of item difficulty values for the MCQ test ......................... 126
Figure 14.2: Results of qualitative analysis of the MCQ Test Feedback document .... 129
Figure 14.3: Spread of mean item values for the Simulation Scenario 1 .............. 132
Figure 14.4: Spread of mean item values for the Simulation Scenario 2 .............. 137
Figure 14.5: Analysis of the simulation scenario feedback ...................................... 142
Figure 16.1: Diagrammatic representation of the domains and dimensions derived from qualitative data on the topic of participant recruitment and logistics .......................................................... 148
Figure 16.2: Map of domains and dimensions derived from qualitative analysis on the topic of simulation assessment execution ............................................................................................................ 151
Figure 18.1: Conceptualisation of the clinical skill sets ........................................... 164
Figure 22.1: ABCDE (Primary Survey) approach to emergency assessment and management ............................................................................................................................................................................ 252
Figure 22.2: Sequential approach to assessment and management in medicine ...... 253
Figure 22.3: Expert approach: Gestalt impression ..................................................... 257
Figure 22.4: Expert approach: parallel patient assessment and management .......... 257
Figure 22.5: Diagrammatic representation of an expert clinician approach to emergency patient assessment and management ...................................................................................................................................................... 258
Figure 22.6: Diagrammatic representation of an expert clinician approach to patient assessment and management in an emergency situation ...................................................................................................................................................... 259
SECTION I: INTRODUCTION

CHAPTER 1: BACKGROUND

A. Introduction

“I will teach them my art without reward or agreement; and I will impart all my acquirement, instructions and whatever I know, to my master’s children, as to my own; and likewise to all my pupils, who shall bind and tie themselves by a professional oath, but to none else.” Hippocratic oath

As doctors, we have a duty not only to our patients but also to our students, who rely on us for their training, and for the feedback they require to better themselves as practicing physicians. The concepts of medical education stretch back for millennia, and are constantly changing. As articulated in the Hippocratic oath, it is our responsibility to provide the best possible education to those who are learning the art and science of medicine, and, therefore, it is our responsibility to ensure that we are using the best possible methods to teach and assess.

As a clinician and an educator, I see that for new medical graduates, despite having been exposed to the clinical environment during medical school, one of the most difficult adjustments is translating their newly acquired medical knowledge into real world clinical practice. This is particularly important in the setting of medical emergencies. The best method of assessing the translation of medical knowledge into clinical practice remains unclear. It seems logical that instead of only assessing knowledge, we should assess practice; the safest and least threatening way to do this may be through the use of simulation.
The intention of this thesis is to contribute to the understanding of the area of
assessment in medical education, in particular the use of simulation as a method for
practical assessment. Specifically, this thesis examines the use of simulation as a tool to
assess the competence of medical students and junior doctors in managing cardiac
emergencies.

This introductory chapter sets the scene for the study: *Validation of Simulation as an
Assessment Tool for Assessing Competence of Medical Students and Junior Doctors in
Managing Cardiac Emergencies* and provides an overview of the rationale of the
research, before ending with the stated research purpose and questions.

The two main focal areas of this study are: firstly, the validation of the simulation
assessment tools that will be utilised in the simulations and, secondly, exploration of the
relationship between this practical form of assessment and the traditional multiple
choice question (MCQ) test method of assessment.
B. Requirements of Junior Doctors

Management of cardiac emergencies is an essential medical skill. The Australian Curriculum Framework (ACF) for Junior Doctors, developed by the Confederation of Postgraduate Medical Education Councils (CPMEC), is “an educational template outlining the learning outcomes required of prevocational doctors”.

The ACF lists the following outcomes related to cardiac emergency management:

1. the ability to assess and manage heart failure, ischaemic heart disease, cardiac arrhythmias, chest pain, breathlessness
2. initiation of resuscitation when clinically indicated
3. identifying when to call for help
4. provision of basic life support and advanced life support

The ACF states that junior doctors are expected to be able to assess and manage cardiovascular conditions, including chest pain, heart failure, ischaemic heart disease and cardiac arrhythmias, “consistent with their level of responsibility”.

Cardiac emergencies, such as acute myocardial infarction (AMI), acute pulmonary oedema (APO) and cardiac arrest, are commonly encountered in the hospital setting, both as presenting complaints in the emergency department, and as secondary issues during admissions.

Acute myocardial infarction (AMI) “is the death of cardiac muscle due to prolonged severe ischaemia”, most commonly caused by coronary arterial occlusion.

Irreversible damage to heart muscle can begin to occur within 20-30 minutes of onset and “half of the deaths associated with acute MI [AMI] occur within 1 hour of onset”.

Immediate management of AMI depends on the clinical situation and
resources available, but may include combinations of aspirin, anticoagulation (heparin or equivalent), oxygen, nitrates, thrombolysis, percutaneous coronary intervention and coronary artery bypass grafting. Without appropriate therapy, in-hospital mortality is approximately 30%, compared to approximately 7% if appropriate treatment is instituted within the recommended timeframe. (2)

Acute pulmonary oedema (APO) is a sudden onset of congestion of the pulmonary vasculature and parenchyma. (2, 3) The pathogenesis of APO is complex and involves multiple systems, but APO is most commonly caused by left-sided heart failure. (2, 3) APO is acutely life-threatening unless immediately and adequately treated. (3) Management may include diuretics, nitrates, morphine, non-invasive ventilation, and importantly, should also include treatment of any underlying cause. (3)

From the first day of internship, junior medical officers are expected to be able to recognise, manage and lead cardiac emergencies with confidence and efficiency. (4-7) While junior doctors are expected to be able to manage cardiac emergencies from the time of graduation, it is uncommon that they would be expected to manage these situations without senior support. In most metropolitan hospitals, there is direct supervision in such situations, such that junior doctors may be responsible for initial assessment and management, but would be expected to seek senior assistance for ongoing management. (1) This concept of clinical escalation is an essential component of good medical practice, particularly for junior doctors, for whom it is crucial to be able to recognise one’s limitations and seek assistance in a timely manner. (1) However, in some cases junior doctors may find themselves in a situation where direct supervision is minimal or not present- for example, community settings, rural rotations
or night duty. It is important that junior doctors are able to manage cardiac emergencies with confidence and competence.(4-7)

In emergency situations such as MET (medical emergency team) calls and Code Blues (cardiac arrest), time is crucial and delays or errors can have a significant negative impact on patient morbidity and mortality.(4, 6) There appears to be little research specifically related to Australian medical education, however overseas studies have demonstrated that medical graduates are underprepared to manage medical emergency situations.(5, 8) It is therefore imperative that we are able to assess the competency of medical students and junior doctors in managing cardiac emergencies for two reasons:

1. To ensure that they are able to provide good quality medical care in cardiac emergency situations,(5, 8) and

2. To provide appropriate further education and training.(4, 8)

This assessment of competence needs to occur without compromising patient safety. It would be inappropriate and unethical to allow a less experienced, less competent junior doctor to manage a real cardiac emergency for the purposes of assessment, and in doing so potentially compromise patient outcomes. It is therefore necessary to provide opportunities for medical students and junior doctors to learn through practice without subjecting patients to any associated risk. Simulation training is a method whereby medical students are able to gain experience in cardiac emergencies, thereby lessening the risk when junior doctors are inevitably exposed to these clinical scenarios.(9-12)
C. Choice of Assessment Methods

Assessment may be used to determine the level of achievement or competence of individuals (or teams) across a range of learning domains and through a progression of achievement levels. (11)

Assessment may be formative, summative or both formative and summative. (13)

Formative assessment is usually undertaken at various stages throughout a course, in order to assess the progress of learners, but with the primary purpose of providing constructive feedback on an individual level, to guide further learning. (11, 13-15)

Formative assessment may also be used to guide or adjust plans for further teaching, depending on the progress of the group. (16)

Summative assessment is generally undertaken at the conclusion of a teaching period (e.g. semester, year, course of study), to determine the overall level of achievement of the learner and often to determine whether or not a learner has passed an assessment and/or can progress to the next stage of learning. (11, 13, 16)

Assessments should be matched to learning objectives that are clearly defined and identified to learners prior to the commencement of the course of study. (11, 16) In cases where assessment is undertaken unrelated to a specific teaching program, for example, for the purposes of (re)credentialing, the objectives of the assessment should still be identified. (16)

In an initial publication in 1956, Bloom identified three domains of learning: the cognitive domain, the psychomotor domain and the affective domain. The cognitive domain relates to all types of learning and assessment, whereas the psychomotor and affective domains are more specific to higher level and applied learning; the
psychomotor and affective domains can be more difficult to objectively assess. (11, 17, 18)

Within the cognitive domain, there were six major categories, which formed what we know as *Bloom’s Taxonomy*. This is a hierarchical framework, which demonstrates the progress of learners through the following stages:(18)

1. Knowledge
2. Comprehension
3. Application
4. Analysis
5. Synthesis
6. Evaluation

A subsequent revised version of Bloom’s taxonomy was later published, which re-defined levels of attainment in the cognitive domain as follows:(18, 19)

1. Remember
2. Understand
3. Apply
4. Analyse
5. Evaluate
6. Create

A further proposed revision of Bloom’s Taxonomy was published by Krathwohl in 2002, which added a second dimension, *The Knowledge Dimension*. This second dimension was applied to each of the six re-defined levels of attainment in the cognitive domain (termed *The Cognitive Process Dimension*), thereby creating a *Taxonomy*
Table, which essentially allows for demonstration of assessment of each level of attainment across different contexts and types of knowledge. Some of these levels of knowledge (factual knowledge, conceptual knowledge and, to some extent, procedural knowledge) may be relevant to both written and practical assessments, though arguably procedural knowledge would be best assessed through means of direct observation of procedural performance. Conversely, metacognitive knowledge requires that the student reflect on his/her cognition, which would usually require completion of some practical task, though the reflective component may be assessable in written form. Krathwohl’s Taxonomy Table is shown in Table 1.1, below (reproduced from Krathwohl’s 2002 paper). According to Krathwohl, “It provides an organizational structure that gives a commonly understood meaning to objectives classified in one of its categories.”

Table 1.1: Krathwohl’s Taxonomy Table (2002)(18)

<table>
<thead>
<tr>
<th>The Knowledge Dimension</th>
<th>The Cognitive Process Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Factual Knowledge</td>
<td></td>
</tr>
<tr>
<td>B. Conceptual Knowledge</td>
<td></td>
</tr>
<tr>
<td>C. Procedural Knowledge</td>
<td></td>
</tr>
<tr>
<td>D. Metacognitive Knowledge</td>
<td></td>
</tr>
</tbody>
</table>
Different authors and researchers have also variously defined levels of competence. Another prominent example is Miller’s Pyramid of Assessment, published in his 1990 sentinel paper entitled *The Assessment of Clinical Skills/Competence/Performance*. Miller’s Pyramid of Assessment defines levels from most basic to advanced as: “knows (knowledge)”, “knows how (competence)”, “shows how (performance)” and “does (action)” (20, 21). Miller’s pyramid was designed specifically in relation to assessment in the medical field, and therefore is more easily applied to the complexity of a field where performance of tasks requires significant theoretical knowledge combined with practical skills and problem solving. Multiple versions of Miller’s pyramid have been (re)produced and (re)designed by various authors. Throughout this thesis, any reference to Miller’s pyramid relates to the version shown in Figure 1.1, which was published by Miller himself in 1990.

![Miller’s Pyramid of Assessment](image)

*Figure 1.1: Miller’s Pyramid of Assessment*
*From Miller, 1990 (21)*
In order to accurately assess attainment of any given learning objective, it is crucial that a suitable assessment method is employed. Logically, then, an assessment method should closely replicate the objective being assessed. Ideally, assessment should involve those who are being assessed actually demonstrating their achievement of the objective.(11)

It is generally accepted that written tests are a suitable form of assessment to assess knowledge on a given topic. Simple recall may be tested by multiple choice question (MCQ) tests or short answer tests. More complex forms of written assessment, such as essay questions, can test higher levels within the cognitive domain, such as comprehension and analysis.(11) However, written tests are not appropriate for testing practical skills; these should be assessed by using a more practical assessment method.(22) Practical assessment methods include objective structured clinical examinations (OSCEs), use of standardised patients, and observational methods.(11)

Miller identified that “the question remains whether what is done in the artificial examination setting ordinarily used to assess any of these elements can accurately predict what a graduate does when functioning independently in clinical practice.”(21)[pS63]. In his sentinel paper, he argued “the most effective substitute for reality is probably the simulated clinical encounter using standardized patients”.(21)[pS65] While this concept has been widely adopted by medical schools in the form of OSCEs, in many situations OSCEs have moved gradually further away from realistic simulation of clinical scenarios, towards an assessment that combines demonstration of a clinical skill with questions, such as those which would be asked in a viva voce (oral) examination.(23, 24)
In addition to OSCEs, there are a variety of other practical assessment methods that are used in medical education. These include:\textit{(11, 25, 26)}

1. Observation of practice
2. Mini-clinical evaluation exercise (mini-CEX)
3. Short- and long-case assessments
4. Direct observation of procedural skills (DOPS)
5. Procedure-based assessment (PBA)

Other methods indirectly assess performance in the clinical setting:\textit{(11, 25)}

1. Portfolios
2. Case-based discussion (CBD)
3. Clinical audit
4. Patient satisfaction surveys
5. Multi-source feedback (MSF), also referred to as 360 degree assessment

Some of the assessment methods listed above are designed for use in assessing those who are already practicing in the clinical setting; they are only appropriate for assessing the progress or ongoing development of a clinician who has already been assessed as competent to work in clinical practice.\textit{(11)} In assessing readiness for practice at the level of students graduating from medical school, it may not be appropriate to simply observe their clinical practice before they are deemed competent and safe to practice. Therefore, at this level, there must be a compromise between an assessment method that reproduces the objective being tested, while also being safe for patients.\textit{(9)}
D. Simulation

Simulation is defined by the Oxford English Dictionary as “The technique of imitating the behaviour or some situation or process … by means of a suitably analogous situation or apparatus”. (27) Simulation has been used extensively in non-medical fields, predominantly for the purposes of skill development and improving safety. For example, the aviation industry uses simulation for training in technical (flight-related) and non-technical (teamwork) skills. Other fields in which simulation has been used include the military, the space programme and the nuclear power industry. (28, 29)

In medicine, clinical simulation involves participants actively performing tasks and/or making decisions as if they were part of a real clinical situation. (30-32) Clinical simulation may take many forms, including role-play, use of simulated patients, low- to high-fidelity simulation and virtual or computer-based simulation. (5, 10-12, 31, 33-36) Simulation allows students and qualified health professionals from a range of fields to learn and refine clinical skills in a safe, non-threatening environment, with room for error and without risk to patients. (4, 6, 7, 10, 12, 28, 33, 37-39)

Clinical simulation for clinical skills training dates back to at least the 1960s, but continues to improve and evolve. (28, 40) Since the 1980s, clinical simulation has also been used for team-based training in medicine. Drawing on the experiences of other industries, the aviation industry’s crew resource management model was adapted to develop the anaesthesia crisis resource management model of training, and military training materials were used in the development of the TeamSTEPPS Patient Safety Program. (28, 29)
Simulation is increasingly used for training in medical education because it is an observed performance method that fits with the principles of adult education, specifically through active participation of the learners, contextualisation and application of knowledge and skills and timely feedback. (7, 10, 11, 41) Simulation is now also being used as an approach in the assessment of clinical skill development. The major theoretical advantages of simulation as an assessment approach include improved reproducibility and the fact that it is more appropriate to assess applied clinical skills using observed performance techniques. (4, 7, 10-12, 42)

E. Differentiating Simulation Assessment from Objective Structured Clinical Examinations

An Objective Structured Clinical Examination (OSCE) is a very commonly used assessment method in medical education, and is most often employed as an end of term summative assessment. (25) An OSCE consists of a series of stations, each of which require the candidate to demonstrate some clinical or practical skill, and may also involve an oral examination component. Occasionally a written component is also employed, though this is less common. (11, 43)

OSCEs are traditionally marked using a checklist assessment tool, though some medical schools use a variation on this, where performance on each “checklist” item, or on components of the station, is graded on a simple three- or five-point scale. (11, 25, 43) OSCEs may also be marked using global rating scales, which, when used by experts, have been shown to have some more favourable psychometric properties, compared to checklist-based assessment (specifically, construct validity, concurrent validity and inter-station reliability). (44) Educators may choose to use a combination of both
checklist and global rating scale for assessment, though a study by Regehr et al demonstrated that the addition of a checklist to a global rating scale did not improve the psychometric properties of an assessment.(44)

OSCEs are a practical mode of assessment, and sometimes use standardised patients or low-to-medium-fidelity simulation.(11, 43) However, the focus in designing OSCEs is generally on creating a reproducible, rather than realistic, assessment.(43) OSCEs are often used in high-stakes assessment, where it is crucial that all candidates’ assessment experience is similar, such that it is a fair assessment.(25, 43) Importantly, OSCEs assess the performance (shows how) level of Miller’s pyramid, rather than the more advanced action (does) level (see Figure 1.1).(11, 21, 42) Therefore, the focus of an OSCE station may be the demonstration of an isolated practical skill, such as suturing or applying a plaster cast.

In contrast to OSCE assessment, simulation-based assessment by definition is a formative assessment, in that a crucial component of simulation is debriefing.(41) This does not mean that simulation cannot be used for high-stakes summative assessment, however, it is necessary that candidates receive immediate feedback as part of the simulation process.(4, 13, 34, 41) Marking of a simulation assessment is very similar to an OSCE: generally a checklist assessment tool, with or without an additional global rating scale or similar, is used to mark candidates.(23, 45) As for OSCE marking, it has been demonstrated that for simulation-based assessment, global rating scales had more favourable psychometric properties (specifically inter-item and inter-station reliability) than checklist-based assessment.(45) Of course, it is crucial that if a global rating scale is used in any form of assessment, that it be criterion-referenced, such that the various levels of performance are well-defined.(46) While these assessment approaches provide
a fairly objective method of assessment,(45) the debriefing component of a simulation also allows for open (though structured) discussion and feedback regarding more subjective observations made by the assessors in relation to the candidate’s professionalism, confidence, fluency, etc.(25, 41)

Unlike OSCEs, the focus of simulation-based assessment is to create a realistic but controlled representation of the clinical environment, where context is important.(10, 34, 47) According to the phenomenon of “context specificity”, student/examinee performance varies depending on the context of assessment (for example, some candidates will perform better in specific clinical contexts), therefore it is important that the context of a simulation-based assessment is well-defined, to allow for standardised assessment of all examinees.(48, 49) Simulation allows for assessment of higher order competencies such as clinical reasoning, rather than only assessing specific practical skills.(26) Therefore, a well-designed simulation assessment should be assessing at least the performance (shows how), if not the action (does) domain of Miller’s pyramid.(21)

F. Evaluation of Assessment Tools

It is important to ensure that assessment tools are correct in their determination of whether or not candidates have achieved any identified objectives. There are two major characteristics of an assessment that should be evaluated, and considered in selection or interpretation of an assessment tool: validity and reliability.(11, 25, 50, 51) It is important to note that validity and reliability measures are evaluated and reported in relation to a particular cohort at a particular time. Psychometric properties of assessment tools, including various measures of validity and reliability, may change when the same assessment tool is applied in a different context, whether that be a different educational
context (different course of study, different point in the same course of study), different clinical context (eg. same assessment tool used to assess a different clinical skillset) or different cohort of examinees.(11, 50) Therefore, it is important that psychometric parameters are constantly re-evaluated in the context of their intended use.(11)

The reliability and validity of an assessment task/tool are particularly important where assessment is *high stakes*. High stakes assessments are commonly used in medical courses and postgraduate vocational training; these are assessments that require a strict passing mark to allow progression to the next level of training and often do not allow for re-assessment.(25)

There are many psychometric measures used to evaluate assessment tools, including:(11, 50-52)

1. Validity
   a) content validity
   b) construct validity
   c) face validity
   d) predictive validity
   e) concurrent validity
   f) formative validity

2. Reliability
   a) interobserver reliability
   b) internal consistency
   c) test-retest reliability
   d) parallel forms reliability
Of these many different measures of validity and reliability, some are most relevant to individual instances of assessment, while others are more relevant to longitudinal progression. This research project sought only to include those psychometric properties that could be evaluated in isolation (i.e. for a single assessment task, rather than in relation to other assessment tasks), at a single timepoint; this was largely due to resource constraints. As such, the particular psychometric properties of interest were: content validity, construct validity, formative validity, interobserver reliability and internal consistency (as part of construct validity). Face validity was not included as this was considered to be less of a strictly psychometric measure, and rather a measure of acceptability.

Each of the psychometric measures used to evaluate assessment tools (including those not evaluated in this research) are defined below:

i. Validity

Validity focuses on what is being assessed, and “is defined as the appropriateness of the interpretations, inferences, and actions that we make based on test scores”. (50)[p172]. Stated more simply, an assessment tool is valid if it measures what it purports to measure. (11, 25, 51-53)

Content validity, also referred to as sampling validity, refers to the relationship between the test items and the objectives of the assessment(51). In general, assessments are conducted to determine candidates’ achievement of a pre-determined set of learning objectives, which should be made available to learners prior to any educational sessions and again prior to assessment. (11, 50, 51) Content validity should be a primary
consideration when designing an assessment tool and is achieved by ensuring that each learning objective is tested by at least one assessment item. However, if an assessment is testing a very large amount of content, it may be impossible to directly test all areas, and therefore it may be necessary for assessment items to sample, but not comprehensively cover, each topic area being assessed.(51)

Construct validity is closely related to content validity, and is defined as “the degree to which an instrument measures the construct it is intended to measure”.(52)[p156] Construct validity is concerned with the influence of other factors on achievement on an assessment task.(51, 52, 54) If particular demographic factors, or knowledge of an area not intentionally assessed by the tool, affected results, then this would result in low construct validity. For example, if an assessment tool were to draw on general knowledge from a particular era, results could be unduly influenced by the age of those being assessed, and this would result in low construct validity. (52) Internal consistency (really a reliability measure, but part of construct validity) refers to the correlation between test items within an assessment, and is a measure of the consistency with which an assessment measures a construct.(50-52)

Face validity is the degree to which the assessment appears to measure what is intended, and is a subjective measure.(11, 51, 52) Contextualisation is an important androgogical (adult learning) principle, and therefore face value may have particular importance in tertiary and continuing education.(15)

Predictive validity is a measure of how well an assessment tool predicts future performance on an independent criterion measure.(50, 52, 54) Predictive validity has been considered particularly important for entry examinations, for example, where the
assessment is intended to determine which candidates are most likely to succeed in a
given domain in the longer term.(50, 55, 56)

Concurrent validity refers to the correlation between scores achieved on the assessment
tool being investigated and another related criterion assessment performed
concurrently.(50, 52, 54, 57) High concurrent validity is desirable when evaluating the
appropriateness of a new assessment tool to replace the one that is currently in use.(54, 57)
Conversely, two or more assessment tools with low concurrent validity may be
intentionally selected to be used together, because in this case, the assessment tools
should test different content or constructs.(57)

Formative validity relates to the information that the assessment tool can provide.(51)
Formative validity may then be thought of in two ways: firstly, the assessment,
especially if used as a formative assessment, may provide useful feedback to candidates
to guide further learning or development;(11, 25) secondly, the assessment may provide
information to educators on areas in which candidates are performing poorly, and
therefore may guide or a different approach to teaching.(51)

**ii. Reliability**

Reliability focuses on the accuracy, stability and reproducibility of an assessment
tool.(11, 25, 50, 51, 53) If an assessment tool is reliable, then the assessment tool will
consistently yield the same result for the same level of performance on the
assessment.(11)

Interobserver reliability, also referred to as interrater reliability or interscorer reliability,
is the correlation between results as determined by two or more independent observers, using the same assessment tool. (50, 51) In any assessment where the assessor must make a judgement on the candidate’s performance, interobserver reliability is very important. (51, 58) Inconsistency between observers may occur due to different interpretation of the marking criteria or different expectations. (51) The degree of subjectivity in assessment using a given tool must be determined and minimized, especially for high stakes assessments; training of assessors may improve interobserver reliability. (51)

Test-retest reliability relates to the temporal stability of an assessment tool. (50-52) To prove test-retest reliability, it is necessary to demonstrate reproducibility of results with the same assessment tool being used to assess the same cohort at different time points. (50-52) Parallel-forms reliability, also referred to as alternative-forms or equivalent-forms reliability, refers to the correlation between results on two different versions of the same assessment tool. (50-52) The alternative assessment tools must be completed by the same cohort of candidates, so that the overall distribution of results, as well as the results of individuals can be examined. (50, 51)

iii. Feasibility

When selecting an assessment method, or developing an assessment, it is also important that reliability and validity are balanced with practical considerations. The feasibility of an assessment tool relates to logistics and resource requirements, and is not so much a characteristic of the assessment tool itself, but rather is situation-specific. (11, 16) Feasibility depends on the financial, physical, time and skill/personnel resources available at a particular organisation or location. (16) Some assessment methods are inherently more or less resource-intensive than others, however some may be initially
resource-intensive to set up and then be able to continue with little ongoing financial investment. (16)

In some cases, it is impractical to use the most accurate assessment method, and therefore some compromise is required to identify an assessment method with a balance of validity, reliability and feasibility. (11) A seminal paper on the topic of Assessment of Professional Competence by van der Vleuten (1996), articulates the relationship between the psychometric and other properties of an assessment tool. The author describes utility as being the product of reliability, validity, educational impact, acceptability and cost. Each of these factors may have carry different weight, but if any of them carries a zero value, then the overall utility of the assessment tool is zero. (59) While the feasibility of the assessment tools is not considered in this research, this should be borne in mind when considering the research findings and their application to educational practice.

G. Objectives
The primary objective of this study was to validate and compare simulation assessments and MCQ tests as methods for assessing competence in cardiac emergency management, in a population of medical students and junior medical officers (JMOs) working at Northern Health (Epping, Victoria).

The secondary objective was to examine the correlation between simulation-based and MCQ assessments of competence in cardiac emergency management. A tertiary objective was identified during the research project in response to difficulties encountered with practical aspects of the research. The tertiary objective was to examine the research process and to identify the strengths, weaknesses, limitations and
issues related to conducting simulation assessment research.

H. Research Questions

The specific research questions were:

1. What are the psychometric properties (reliability and validity) of an MCQ test and a simulation assessment, both designed to assess management of common cardiac emergencies?

A) What is the construct validity of each method of assessment?
To what extent does each method of assessment actually measure the construct of interest, that is, participants’ competence in managing cardiac emergencies?(51)

B) What is the formative validity of each method of assessment?
To what extent does each method of assessment provide information that can be used to inform further participant training/education in management of cardiac emergencies?(51)

C) What is the internal consistency (specifically inter-item correlation) of each method of assessment?
What is the correlation between scores within each assessment tool?(51)

D) What is the inter-observer reliability of cardiac emergency simulation as an assessment tool?
To what extent does scoring on a simulation-based assessment of cardiac emergency management vary between assessors?(51)

E) What is the content validity of each method of assessment?
To what extent does each method of assessment evaluate achievement of the relevant learning objectives, as documented in the ACF?(51)
2. What is the relationship between performance on simulation-based and MCQ assessments of cardiac emergency management?

What is the correlation between participants’ scores on a simulation-based assessment and an MCQ test both designed to test cardiac emergency management?

3. How do the psychometric properties of the MCQ test and the simulation assessments compare?

What are the differences in the measures of reliability and validity obtained for the two assessment methods?

4. What are the strengths, weaknesses, limitations and issues related to conducting simulation assessment research?

I. Summary and Thesis Structure

*Chapter 1* of this thesis has given some background on the issues surrounding the topic of using simulation to assess competence of medical students and junior doctors in managing cardiac emergencies. Graduating doctors are expected to be able to recognise, manage and lead cardiac emergencies (4-7) and, therefore, medical educators must be able to assess the competency of medical students and junior doctors in managing cardiac emergencies.(5, 8) There are many assessment methods available to medical educators, each with advantages and disadvantages, which must be evaluated in relation to the specific assessment situation and aims.(11, 16, 22) The purpose of this research is to evaluate and compare the traditional MCQ testing method and the newer simulation assessment method, as approaches for assessing competence in managing cardiac emergencies.
Chapter 2: Literature Review discusses the currently available research and evidence on simulation as an assessment method and cardiac emergency simulation, and demonstrates the need for further research.

Section II of the thesis, entitled Methodology, gives an overview of the methods used for this research (Chapter 3: Methods Overview) and describes the development of the multiple choice question (MCQ) and simulation assessment tools (Chapter 4: Assessment Tool Development). Chapter 5 describes the methods used to analyse participant characteristics and overall results. Chapters 6-9 describe, in detail, the specific statistical and qualitative analysis methods used to examine each of the psychometric properties. Chapter 10 describes the qualitative methods used to examine the research process.

Section III (Chapters 11-16) of the thesis presents the Results of the research. Again, separate chapters present findings for participant characteristics and overall results, each of the psychometric properties, and the research process.

Section IV (Chapters 17-22) discusses the results and their implications, in terms of addressing the research questions.

Section V concludes the thesis, providing a Summary of the Research Findings (Chapter 23), Recommendations (Chapter 24), a discussion of the Limitations of the research (Chapter 25) and possible Future Directions (Chapter 26). Chapter 27: Conclusion concludes this thesis.
CHAPTER 2: LITERATURE REVIEW

A. Introduction

Chapter 1 has provided a background to the requirements of junior doctors and assessment in medical education in general, and stated the objectives of the study. The purpose of the literature review presented in this chapter is to identify and present the current available research and evidence relating to the preparedness of medical graduates to manage cardiac emergency situations, methods for improving and assessing graduates’ preparedness for such situations, the use of simulation in teaching and assessment, and development of simulation assessment tools. Each of these is discussed in the subsections of this chapter.

Potentially relevant literature was identified using both the UWA OneSearch Tool (which searches across a range of physical reference materials and online materials held by the University of Western Australia, including journals, e-books and databases), and Google Scholar. Keywords used were: simulation, assessment, emergency, cardiac, medical. Journal articles and other published reports/papers that were identified by the searches are included in the literature review for their relevance to either one or both of:

1. Use of simulation in teaching or assessment of students or graduates in medical, nursing or allied health fields
2. Preparedness of medical graduates to manage cardiac emergencies.
B. Medical Graduates’ Preparedness to Manage Emergency Situations

Adequate training of medical graduates to be able to manage emergencies is crucial to patient safety/outcomes. (5, 6) There is no available literature that discusses Australian medical graduates’ readiness to manage emergency situation. However, a recent study by Karakus et al, published in 2014, demonstrated that final year medical students in Turkey were underprepared to manage emergency situations, including gastric perforation, major blood loss secondary to trauma, aortic dissection and deep vein thrombosis with pulmonary embolism. (5) It is unclear if these findings would be generalisable to medical graduates in other countries, specifically Australian medical graduates.

It is necessary for junior doctors to be able to manage emergency situations confidently from the time they graduate, as they provide initial assessment in emergency cases, and are often the only doctors present initially in the case of patient deterioration. (4, 8) In these situations, doctors must take an organised approach to provide efficient and effective care. (4, 6) Karakus et al also propose that improved training would also be expected to decrease healthcare expenditure and litigation. (5)

Various approaches have been suggested to improve medical graduates’ preparedness to manage emergency situations. (5) Sahu and Lata (2010) assert that “Traditional methods of educating residents and medical students using lectures and bedside teaching are no longer sufficient…It is unreasonable to expect the educational model developed 50 years ago to be able to adequately train the medical students and residents of today.” There is also increasing recognition that andragogical principles need to be incorporated into medical courses, acknowledging that adult learners have different needs to younger learners. (10) Simulation is one way in which medical courses are addressing this
C. Advantages of, and Barriers to, Use of Simulation as a Teaching Method in Medical Education

Simulation is increasingly being used in undergraduate medical training to achieve a range of learning outcomes. Many authors argue that simulation-based training provides advantages over traditional classroom-based training, especially in the context of emergency management. In 2010, McGaghie and colleagues published *A critical review of simulation-based medical education research*, which discusses the current state of this educational approach, and proposes 12 features and best practices for simulation-based medical education. The key findings of this review, and other relevant literature, are discussed in this section.

Firstly, simulation allows students the opportunity to practice practical patient management, sometimes managing situations that they would not otherwise be involved in, in a safe and controlled environment, where mistakes will not compromise patient safety. While the patients are protected from harm, the participants experience very real feelings of pressure and anxiety, which can help prepare them for dealing with these emotions in real life.

Repeated exposure to simulations can lead to habituation, a process where stress levels decrease as participants grow accustomed to the simulation environment and process. Habituation may be beneficial, in that simulation does not cause ongoing high stress for participants. Repeated practice in a simulated environment can also help train participants to multi-task and manage distractions and stress that may be encountered in
clinical practice. However, habituation also means that the novelty and impact of simulation fades, perhaps lessening both the appeal and the benefits over time.\(^\text{(63, 64)}\)

Overall, exposure to simulated clinical settings better prepares students for patient care, compared to classroom-based teaching.\(^\text{(7)}\) Several studies have proven transfer of simulation-based training to patient care settings. Training in specific skills, whether pure procedural or complex management skills, using simulation, results in improved performance on those skills in practice.\(^\text{(60)}\)

According to the literature, other benefits of simulation-based training are that this approach is enjoyable for participants, it contextualises the teaching, and encourages self-reflection.\(^\text{(4, 10, 65)}\) Simulation-based training also helps to build self-confidence and develop decision-making processes.\(^\text{(37, 65)}\) When simulations are conducted in teams, this also helps to develop teamwork skills.\(^\text{(7, 60)}\) As a practical approach, simulation is particularly useful for clinical skills development but, especially in conjunction with some teaching on theory, is also able to be used to teach knowledge-based content.\(^\text{(4, 61, 65)}\) Simulation-based training is able to be used as an approach for almost any specialty area.\(^\text{(4)}\)

A major argument for use of simulation-based training in medical education, articulated in a review article by Sahu and Lata, is that this approach best addresses adult learning principles, by allowing for application of what is being learnt, and building on experience.\(^\text{(10, 65)}\)

In implementing simulation in medical education, it is important to consider potential barriers, so that these may be addressed. Heitz et al (2011) conducted a survey-based
study of medical educators, which identified several perceived barriers to instituting simulation-based training for medical students. These included faculty time/availability, time available in the medical education program, number of students, issues relating to costs/funding and faculty training. (12) Most of these barriers relate to resource availability and allocation. The authors of the study suggested that some of these barriers might be overcome by replacing existing material in the medical curriculum with simulation-based modules, where appropriate, and using hybrid approaches to combine simulation-based and classroom-based teaching. (12) It is also important to recognise that, while there is a significant outlay of resources required to implement simulation-based teaching in the initial stages, this is an investment in a program that will become largely self-sustaining once established. (36, 38, 59, 66)

Another significant barrier for implementation of simulation in medical education is a sense of inertia or concern/trepidation at the concept of replacing traditional methods with newer methods, and the cultural change required to accept simulation into mainstream medical education. Educators and students who are familiar with conventional teaching methods may prefer that which is familiar and be reluctant to accept change; this has also been a barrier to implementation of interprofessional education in medical education. (35, 38, 60, 67) As previously mentioned, however, it is not necessary to discard traditional methods in order to implement simulation; simulation can be combined with classroom-based approaches. (12) In fact, it is recommended that simulation-based teaching should be integrated into the medical education curriculum with other teaching approaches. (60)

One of the most commonly cited barriers to use of simulation in medical education is that it is very resource-intensive, requiring significant time, personnel and funding. (41)
Computer-based simulation has been suggested as a feasible alternative teaching method to live simulation, with a major advantage associated with this approach being economic efficiency. (5, 33) This approach is commonly used in some countries, and was demonstrated, in a paper by Karakus et al (2014), to improve student knowledge of management of common emergencies. (5)

Possible disadvantages of computer-based and virtual simulations include the human resource requirements to create computer-based simulations (as well as involvement of clinicians and educators, personnel with computer programming skills are required), and that in designing the simulations there is a tendency to focus on use of technology, rather than teaching principles. (33) Of course, the other obvious disadvantage of computer-based simulation is the lack of realism, in terms of lack of authenticity and lack of interaction with a real patient; however, the authenticity of computer/virtual simulations is improving with the use of new technology. (68) There is also an inherent lack of realism in the use of algorithms for virtual simulations that require trainees to select from pre-defined choices, and that are linear in their design. (33)

Another alternative to address resource issues was investigated by Cooper et al, and involved using residents, rather than faculty, to facilitate simulation training. (41) This study showed that residents were perceived by medical students to be as effective as faculty in simulation-based teaching, and the authors went on to recommend that residents may be used as facilitators to increase the capabilities of a department to use simulation in teaching. (41) Review of the literature did not identify any studies addressing resource issues for simulation-based assessment.
D. Comparison of Simulation with Other Teaching Methods

Ruesseler et al (2010), recognising the lack of published research comparing simulation training to other teaching methods for improvement of practical emergency medicine skills, conducted a study in which they used simulation-based training to teach final year medical students on topics including basic life support, advanced cardiac life support, advanced trauma life support and common emergencies. This study demonstrated that simulation was an effective method for teaching management of emergency situations, and that the group of students who were taught using this method achieved significantly better assessment results (on OSCEs) than those who were exposed to traditional teaching in the emergency department. As a result of this study, simulation-based training was incorporated into the standard teaching curriculum at Frankfurt Medical School, Germany.

Unfortunately, this study had a very small sample size (22 students in each of the simulation-based teaching and traditional teaching groups), and it is unclear whether the findings are generalisable to teaching different topics and students of different levels of experience. The other issue with this study is that the primary outcome measure was performance on OSCEs, so it is possible that the improved performance by the group who were taught using the simulation-based approach is related to increased confidence and exam technique in performance-based assessment. It would be pertinent to demonstrate that this improved performance on OSCEs transfers to improved clinical practice, as has been demonstrated in other studies on simulation-based education.

Another study conducted by Steadman et al has compared the use of simulation-based training with problem-based learning (PBL), and in this study, simulation was used as the assessment method to determine the difference in performance between the two
groups. The major finding for this study was that “For fourth-year medical students, simulation-based learning was superior to problem-based learning for the acquisition of critical assessment and management skills.”(69) As with Ruesseler’s study, the method of assessment used in Steadman’s study also calls into question whether the difference demonstrated is due to the simulation group having more confidence and experience with practical assessment, rather than necessarily demonstrating a true improvement in clinical competence. It is also important to note that this study only demonstrates acquisition of skills, with participants re-tested on day 5 of the educational program, and does not investigate longer-term retention of these skills.(69)

E. Assessing Performance of Examinees on Simulations

The studies examined in this literature review show that the methods used to assess examinee performance on simulation assessments are quite varied, and that there is no consensus as to the best approach. As such, the specific methods used by different studies are of interest.

In 2010, Luscher et al published the results of their study “Proficiency in cardiopulmonary resuscitation of medical students at graduation: a simulator-based comparison with general practitioners”. This study focused on outcomes related to proficiency in cardiac arrest management.(8) In this study, simulations were recorded and assessed at a later time by two different assessors separately, with subsequent discussion between the assessors if there were any disagreements. The interobserver reliability prior to discussion, as measured by Cohen’s Kappa, was excellent (>0.9). Assessment involved some timed measures, such as “hands-on time” (cardiac massage/defibrillation) and time to “first appropriate intervention”. A checklist was also used to assess behavioural ratings and to assess the quality of resuscitation measures.(8)
As previously mentioned, a study by Steadman et al used simulation to assess acquisition of critical assessment and management skills. This study used a weighted checklist assessment tool (items were weighted based on their importance) to score performance on initial and final simulation assessments. (69)

The simulation assessment methods used in those studies focusing on evaluation of the psychometric properties of simulation assessment tools are discussed further in Chapter 21: Studies Evaluating Psychometric Properties of Simulation Assessments, and are then summarised in Table 2.2.

F. Benefits of Simulation as an Assessment Method

One of the benefits of practical assessment methods, such as simulation, is that content being assessed is contextualised. This adds face validity/relevance for students and these types of assessment also encourage self-reflection. (4) Perhaps, though, the major benefit of using simulation for assessment is that, when highly authentic (high fidelity) simulation is used, the testing situations closely approximate the real life situations for which the examinee’s competence is being assessed. (59, 60)

Luscher et al advocate for simulation as a method to assess medical emergency management. They argue that simulation provides reproducible clinical scenarios and avoids the ethical and practical issues around allowing students to manage real emergency situations. With simulation, risk to patients is avoided, and competence can be fully assessed, because it is not necessary to intervene if a student’s performance is poor or potentially dangerous. (8)
Luscher et al also claim that simulation has the potential to fulfil all the requirements to assess at the highest level of Miller’s pyramid, that is, *action*. This is in line with Miller’s own argument that simulated clinical encounters provide the closest approximation to reality. (21) However, while simulation provides the closest *approximation* to reality, it is important to emphasise that simulations are performances of what examinees know they *should do*, rather than an observation of what they *actually do* in practice. As such, this would correspond with the ‘performance’ (second highest) level of Miller’s pyramid, rather than the ‘action’ (highest) level, which can only be assessed by observation of real practice. (11)

G. Simulation in Teaching and Assessing Cardiac Emergency Management

Cardiac emergency simulation may take many forms, including standardised patients, CPR simulators, the *Harvey* cardiology simulator, high-fidelity mannequin simulators (including the most modern and realistic human patient simulator models) and procedural simulators (for more invasive procedures). (10) In 2010, Sahu and Lata published a review article relating specifically to use of simulation in resuscitation training. In this review, cardiac emergency management was identified as a specific area “needing a fundamental shift in the teaching model”. The authors argue that use of high-fidelity mannequin simulation in teaching and assessment, can ensure that graduating doctors are able to manage cardiac emergencies. (10)

There has been very little published research on use of simulation as a method for teaching or assessing cardiac emergency management outside the very specific domain of cardiac arrest/immediate resuscitation management (basic and advanced life support). Some studies have incorporated cardiac scenarios into a wider simulation education or assessment program, (42, 70) and other studies have addressed resuscitation...
management,(42) but there has been little research into the use of simulation specifically in medical management of cardiac emergencies, outside the context of immediate resuscitation.

The only study identified that looked specifically at using simulation to assess medical trainees’ management of cardiac emergencies was a 2010 paper by Opar et al., published in the *Journal of Graduate Medical Education*. In this US study, the participants, who were PGY-1 residents (equivalent to our intern year) were assessed on their performance in a single simulation scenario, which depicted a patient presenting with acute coronary syndrome, who then deteriorated to cardiac arrest. The study was longitudinal, assessing resident performance before and after completion of their internship, with the primary outcome being improvement in performance after internship. The primary outcome was achieved, with improvement on three of the four measures (chest pain score, medical knowledge performance, and systems-based practice performance), though, interestingly, a decline in patient care performance was demonstrated.(71) The validity of the assessment tool was assumed, rather than being specifically evaluated. While this is not ideal in terms of educational practice,(50) this does demonstrate a move towards acceptance of simulation-based assessment in medical education.

**H. Studies Evaluating Psychometric Properties of Simulation Assessment Tools**

It is recommended that assessment tools be evaluated to ensure that they are both reliable and valid.(50) Other important factors in determining the utility of an assessment tool depend on the individual educational context, and include educational impact, acceptability and cost; these factors are not considered in this section, since they
are not psychometric properties of assessment tools and generally not the subject of educational research. (59) The aim of this section of the literature review was to examine the way in which different studies have approached evaluation of simulation assessment tools, and their findings, to assist in design of this research project. Review of the literature identified only the seven studies below that evaluated simulation assessment tools in medical or nursing education, with most studies instead focussing on simulation design, or use of simulation for interprofessional education.

Cazzell and Howe conducted a study to examine interrater reliability using a checklist assessment tool for assessing performance on a paediatric medication administration simulation undertaken by nursing students. This study used kappa values and intraclass correlation coefficients to measure the interobserver reliability for individual checklist items. Results demonstrated mostly moderate agreement between assessors based on kappa values, and poor-to-moderate reliability based on intraclass correlation coefficient values. The study concluded that there was difficulty in measuring some behaviours using the checklist and that this reinforces “the need for consistency in rater roles.” (72)

The major strengths of the study by Cazzell and Howe were the sample size of 207 students, the use of more than one measure of interrater reliability, and the analysis of interrater reliability by learning domains. The major limitations were lack of assessor training in use of the assessment tool, and a lack of discussion on two items deemed to be “not reliable” (not include in those described as “unacceptable interrater reliability”). This study also examined only one psychometric property (interrater reliability), therefore the implications of the findings are limited, as the broader reliability/validity of the assessment tool is not addressed. While this study relates to nursing students,
results should be generalisable to OSCE/simulation-type assessments in other medical and allied health fields, as long as a similar assessment tool is used.\(^{(72)}\)

Foell et al conducted a “validation study of the da Vinci Skills Simulator” (dVSS), which evaluated simulation-based assessments using a surgical skills simulator. The authors evaluated face and content validity together, and their paper did not explicitly define the methods used to evaluate these properties, although it may be assumed from the reporting of the results that a questionnaire tool was used. They reported excellent face and content validity for use of dVSS as an assessment method. Construct validity was evaluated using a contrasted-groups approach, and the assessment method was evaluated as having demonstrated “excellent” construct validity, since expert participants performed significantly better than novice participants on five of seven exercises. Concurrent validity was evaluated by comparing participants’ performance on the seven different assessments (all conducted using the dVSS), to their performance on two robotic surgery exercises. Concurrent validity was reported as being excellent, however this was based on p-values obtained, rather than the magnitude of the Pearson correlations.\(^{(73)}\)

The major strengths of the study by Foell et al were the use of an objective scoring system, which was built into the robotic surgical simulator, the assessment of construct validity using a contrasted-groups approach, and the fact that multiple areas of validity were assessed. The limitations of the study lie in the lack of clarity regarding methods used to assess face and content validity and the relatively small sample size. While this study is relevant to the topic of assessing competence of junior medical trainees, the assessment in this study is of a skill set that is less broad than cardiac emergency management, and is able to be more objectively assessed.\(^{(73)}\)
A study by Gupta et al assessed interrater reliability of a trauma simulation assessment tool. This was a very small study, with only eleven participants assessed by four assessors each. The assessment tool was not a standard checklist, but rather performance on five core competencies was marked on a scale of 1-5. Fleiss’ kappa was used to measure interobserver reliability, since there were multiple assessors for each instance of assessment. The study found low interrater agreement based on this measure. This was a very small study and as such, the implications and generalisability of the findings are very limited. (74)

Lypson et al published a study evaluating an assessment tool for aseptic technique in resident physicians. The assessment task was a simulation that required participants to set up for, and pack up after, a sterile procedure. The assessment tool included a checklist component and global rating score. The researchers used a contrasted-groups/comparison across all groups approach as well as Cronbach’s alpha to evaluate construct validity and internal consistency. They also used intraclass correlation coefficient to evaluate interobserver reliability. The main findings of the study were that there was adequate internal consistency, interobserver reliability and construct validity. However, they also found that this assessment tool was unable to distinguish between performance for residents of more similar levels (postgraduate year 1 residents, compared to postgraduate year 2/3 residents). (75)

The major strength of the study by Lypson et al was the evaluation of multiple psychometric properties. The limitations of the study were that not all checklist items were included in statistical analysis, which may bring into question the validity of the results reported, and the small number of participants in the second part of the study (validation study). Further studies using a similar approach with larger numbers, and
including all checklist items in evaluation of psychometric measures would be useful, and generalisable to other simulation assessment tools. (75)

Reid et al published their research on the “Simulation Team Assessment Tool (STAT)” in 2012, which discussed the development, reliability and validation of this assessment tool, designed to assess team performance in simulated paediatric emergencies. This study included evaluation of interrater reliability using intraclass correlation coefficients and evaluation of construct validity using a contrasted-groups approach. The study found that the assessment tool had good to excellent interrater reliability and adequate construct validity. However, the study was significantly limited by sample size, with only four teams assessed. As such, the results of the study cannot be generalised outside the single administration of this simulation assessment. (76)

Neira et al also published research that discussed the development and validation of another simulation assessment tool, the “Generic Integrated Objective Structured Assessment Tool (GIOSTAT)”, which was used to assess management of anaesthetics emergencies (two different scenarios). In this case, interrater reliability was also calculated using intra-class correlation coefficients. Construct validity was investigated by correlating assessment scores with residents’ level of experience. This study found that interrater reliability was substantial, and that there was significant correlation between assessment performance and experience, indicating construct validity. (70)

The validity of the study design used by Neira et al is questionable; a retrospective approach was taken, in which video recordings of participants partaking in simulations were assessed. Some of the recordings had been used in the pilot study, which was part of the development of the assessment tool, and it was unclear whether or not these participants were randomised to the two different scenarios (though this seems unlikely...
given the uneven number of participants completing each scenario). Other recordings that were assessed were simulations that were not initially intended to be marked by the assessment tool and had been undertaken at a earlier time, with participants having been randomised to one of two scenarios, which were the same as the scenarios used for the pilot study. While the results of the study by Neira et al may give some information on the validity of the assessment tool used, a prospective study with a larger number of participants randomised to each of the two scenarios, would give more valid and generalisable results.\(^{(70)}\)

The article by Hall et al describes the development and validation of the “Queen’s Simulation Assessment Tool” (QSAT), designed to assess performance on resuscitation simulations. The study was conducted over a period of two years, and involved emergency medicine postgraduate trainees at different levels of experience. Ten different simulation scenarios were included in the study, with each participant completing two or three scenarios; allocation to scenario depended on when the participants were being examined, as simulation scenarios changed each semester. Validation of the assessment tool was based on interrater reliability, construct validity and acceptability to participants. It should be noted that the authors used interclass correlation coefficient to determine interrater reliability, in contrast to other studies that used \textit{intraclass} correlation. The researchers performed generalisability studies to evaluate for contributors to variance in results.\(^{(42)}\)

The study by Hall et al showed that the QSAT showed adequate interobserver reliability. Construct validity, as determined by a contrasted-groups approach, was strong. Generalisability studies found that the largest contributors to variance were, appropriately, trainee, and trainee by scenario, which confirms construct validity.
Results of the participant questionnaire demonstrated that the assessment was acceptable to participants and perceived by them as being valuable to learning.\(^{(42)}\)

While the overall number of participants was high, there were only 19 to 25 residents completing each simulation scenario, such that there is not a large amount of data on reliability and validity of the assessment tool for each individual simulation scenario.\(^{(42)}\)

Based on the seven studies discussed above, the most commonly evaluated psychometric properties and the statistical methods used to measure these properties are shown in Table 2.1, below.

**Table 2.1: Most commonly evaluated psychometric properties and methods used to measure these properties**

\(^{(42, 70, 72-76)}\)

<table>
<thead>
<tr>
<th>Psychometric Property</th>
<th>Method of Evaluation</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interobserver reliability</td>
<td>Intraclass correlation coefficient</td>
<td>Cazzell and Howe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lypson et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reid et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neira et al</td>
</tr>
<tr>
<td></td>
<td>Interclass correlation coefficient</td>
<td>Hall et al</td>
</tr>
<tr>
<td></td>
<td>Kappa (Cohen’s or Fleiss’)</td>
<td>Gupta et al</td>
</tr>
<tr>
<td>Face and content validity</td>
<td>Questionnaire</td>
<td>Foell et al</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Contrasted-groups approach</td>
<td>Foell et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lypson et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reid et al</td>
</tr>
<tr>
<td></td>
<td>Correlation with level of experience</td>
<td>Neira et al</td>
</tr>
<tr>
<td></td>
<td>Internal consistency</td>
<td>Lypson et al</td>
</tr>
<tr>
<td></td>
<td>(Cronbach’s alpha)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Generalisability study</td>
<td>Hall et al</td>
</tr>
<tr>
<td>Concurrent validity</td>
<td>Comparison to performance on set exercises</td>
<td>Foell et al</td>
</tr>
</tbody>
</table>

Table 2.2, next page, summarises the research methods and major findings of the studies evaluating psychometric properties of simulation assessment tools.
<table>
<thead>
<tr>
<th>Study</th>
<th>Research Design</th>
<th>Simulation Scenario(s)</th>
<th>Population(s)</th>
<th>Sample Size</th>
<th>Simulation Assessment Tool</th>
<th>Outcome Measure(s)</th>
<th>Main Findings</th>
</tr>
</thead>
</table>
| Cazzell and Howe (2012) | Prospective observational | Medication administration | Nursing students               | 207         | 14 item dichotomous checklist (12 items used for interrater reliability) | Interrater reliability                                                           | - adequate interrater reliability on 6/12 items (cognitive & psychomotor domains)  
|                |                          |                        |                                |             |                           |                                                                                | - inadequate interrater reliability on 4/12 items (affective domain)             |
| Foell et al (2013)  | Prospective observational | Robotic surgery skills | Surgical residents, fellows, staff surgeons | 53          | Built in scoring algorithm, 7 questionnaire | Face validity                                                                 | - excellent face validity                                                            |
|                |                          |                        |                                |             |                           |                                                                                | - excellent content validity                                                       |
|                |                          |                        |                                |             |                           |                                                                                | - excellent construct validity                                                     |
|                |                          |                        |                                |             |                           |                                                                                | - excellent concurrent validity                                                   |
| Gupta et al (2010) | Prospective observational | Trauma management      | Emergency medicine residents   | 11          | 5 competences marked on 5 point scale | Interrater reliability                                                           | - low interrater reliability                                                      |
| Lypson et al (2010) | Prospective observational | Aseptic technique      | Medical students, surgical residents, theatre nurses | 331         | 20 items marked on 3 point scale (13 items used for psychometrics); global 9 point rating scale | Internal consistency, Interrater reliability, Construct validity                | - moderate to high internal consistency                                           |
|                |                          |                        |                                |             |                           |                                                                                | - variable interrater reliability for checklist items (poor to high)             |
|                |                          |                        |                                |             |                           |                                                                                | - good interrater reliability for global rating scale                            |
|                |                          |                        |                                |             |                           |                                                                                | - variable (on different measures) but adequate construct validity               |
| Reid et al (2012)  | Prospective observational | Paediatric resuscitation | Paediatric residents, resuscitation experts | 4 teams     | 94 items marked on 3 point scale | Interrater reliability, Construct validity, Content validity                  | - good to excellent interrater reliability                                       |
|                |                          |                        |                                |             |                           |                                                                                | - good construct validity                                                        |
|                |                          |                        |                                |             |                           |                                                                                | - content validity addressed during design                                        |
| Neira et al (2013) | Retrospective observational | Anaesthesitics emergencies | Anaesthesics residents      | 50          | 14 items marked on 6 point scale | Interrater reliability, Construct validity                                   | - “substantial” interrater reliability                                          |
|                |                          |                        |                                |             |                           |                                                                                | - variable construct validity (significant for one set of competencies, no correlation for the other) |
| Hall et al (2015)  | Prospective observational | Resuscitation          | Emergency medicine postgraduate trainees | 92          | 4 domains marked on 5 point scale; overall score marked on 5 point scale; questionnaire | Interrater reliability, Construct validity, Acceptability (to participants) | - adequate interrater reliability                                                 |
|                |                          |                        |                                |             |                           |                                                                                | - strong construct validity                                                       |
|                |                          |                        |                                |             |                           |                                                                                | - acceptable to participants, perceived as valuable to learning                  |

Table 2.2: Research methods and major findings of the studies evaluating psychometric properties of simulation assessment tools

Anastasia Vlachadis Castles
21544591
I. Studies Describing Simulation Assessment Tool Development

There were only limited studies that discussed the development of the simulation assessment tools they used, or even described the assessment tools in detail. While some other studies described the development of simulation scenarios, most of these related to use of simulation as a teaching method, rather than as an assessment, and therefore did not include development of an assessment tool.

The study by Cazzell and Howe (discussed earlier in this chapter) only very briefly described the development of their assessment tool, used to assess paediatric drug administration by nursing students. The authors simply stated “The OSCE checklist was developed by a team of paediatric nurse educators to reflect the step-by-step processes (knowledge, skills, and professional communication) of safe medication administration taught throughout the undergraduate nursing curriculum.”[72][p221] (Note: the assessment was termed an OSCE but was run as a clinical simulation, and included a debrief component.) The published article included the complete assessment checklist, which was a simply structured list of “steps of proper medication administration” and assessed in a binary fashion (yes/no).[72]

Lypson et al, also did not provide much description of the process of developing their assessment tool for aseptic technique, but published the complete assessment tool. It was noted that the checklist component of the tool was not marked in a binary fashion, but rather the assessors could select from three levels of achievement for each item: done/ needs improvement/ not done. There was also a global rating component of the assessment tool, where the participant’s overall performance was rated on a scale from one (defined as novice) to nine (defined as master). Lypson et al also reported that some items were excluded from the final scores, but did not discuss the reasons for this in
Reid et al described the process for developing their Simulation Team Assessment Tool (STAT), as a tool to assess team management of any paediatric resuscitation. The process was described in some detail; it included derivation from pre-existing assessment tools, determination of the final items using the modified-Delphi method, expert review, editing, and consensus on items and scoring. Again, the assessment tool was published: it was structured as a checklist with a three-point scale for most items (complete and timely/ incomplete or untimely/ needed and not done), and an option to mark items as “not required”.

The process followed by Neira et al in their design of an assessment tool for anaesthetic emergencies, was extremely comprehensive, and made up a large part of the research. Essentially, the researchers used a process that involved literature review, item writing, pilot testing, re-evaluation by experts, modification, and implementation. Again, the modified-Delphi method was used in this process. In this case, the assessment tool was not included in the published paper, but is described as a two-part tool with fourteen items overall, each marked on a six-point scale, with a description provided for each point on the scale.

Hall et al used a similar approach to Reid et al in developing their QSAT. The development process included multiple rounds of expert review, and utilised a modified-Delphi method. The full assessment tool (generic version) was included in the publication, and included assessment on four domains (primary assessment, diagnostic actions, therapeutic actions, communication), as well as overall performance. Each of
these was scored on a five-point scale, with descriptions given for each possible mark. (42)

The processes and format of assessment tools used in the articles above, along with other resources, were used to guide development of the simulation assessment tools for this research project, and the evaluation of the assessment tools. Specifically, the simulation assessment tools were based on the checklist formats used by Cazzel and Howe, Lypson et al, and Reid et al. (72, 75, 76) The methods of evaluating psychometric properties were based on those shown in Table 3.1; however, where there were multiple different evaluation methods used by different researchers or other alternative methods available, the approach(es) that best matched the data type(s) was used. The development processes for both the MCQ test and simulation assessments, and the choice of psychometric evaluation methods, are discussed later in this thesis.

J. Conclusions

This literature review has demonstrated that, according to available research, medical graduates are underprepared to manage the emergency situations that they will likely encounter and be expected to manage as junior doctors. (4-6, 8) The advantages and barriers to use of simulation in medical education have been discussed, as have methods to overcome the barriers. The many uses and benefits of simulation in assessment have also been outlined.

This review process has identified three major areas where there is paucity in available literature. There is a lack of literature available discussing or describing methods to assess content validity or formative validity of simulation-based assessments.
specifically. There was only one published study identified that related specifically to cardiac emergency simulation for assessment, but this study did not validate the assessment tool in terms of psychometric properties.

Importantly, the literature review has informed the design of the research project. The research methods used in previous studies evaluating the psychometric properties of simulation (Table 2.2) were used to guide the overall research methodology for this research process. Attempts were made to avoid deficiencies identified in previous studies (e.g. small sample size, retrospective approach) and the approaches that were successful were emulated in the study design for this project (e.g., prospective approach, evaluation of multiple psychometric properties).

The specific methods used to evaluate psychometric properties in previous studies (summarised in Table 2.1) were used to guide the methodology and statistical analysis methods used to evaluate interobserver reliability and construct validity in this study. Based on the paucity of previous research describing evaluation of content validity and formative validity, a decision was made to develop new methodologies for the evaluation of these psychometric properties.

The approaches to simulation assessment tool development previously used by other researchers were used to guide assessment tool development for this project (e.g. expert review, piloting). Similarly, the descriptions and examples given of the simulation scenarios and assessment tools used in previous studies (clinical scenarios, checklist structures, etc.) were used to guide the format and structure of the simulation scenarios and assessment sheets for this project.
K. Summary

This chapter has presented a summary of the current available research and evidence relating to simulation as an assessment method. The literature review also revealed areas in which there was a lack of previous research, which will be addressed in this thesis. This literature has influenced the design of this research project, specifically the simulation assessment tool design and the research methodology. The following section will describe the methodology used for this research project.
SECTION II: METHODOLOGY

CHAPTER 3: METHODS OVERVIEW

This chapter describes the methods used in this research project, including the general approach, the research process, research ethics, participant recruitment, data collection, and a brief overview of analysis methods. Details of statistical and qualitative analysis methods used are included in the Chapters 5-10.

A. Research Approach

This research project was a prospective, mixed methods, experimental, descriptive and correlational study.(50) The study was experimental, rather than observational, in that an artificial investigational situation was created, in which participants were divided into groups for comparison, and completed assessments specifically designed for this research project; that is, it was not part of their usual study- or work-related assessment. Participants’ results on the assessment tasks were used to descriptive and inferential statistics, and performance on the two assessment methods (MCQ test and simulation assessment) was correlated.

A mixed methods (combined quantitative and qualitative) approach was used for this study. Greene et al, in an article entitled “Toward a Conceptual Framework for Mixed-Method Evaluation Designs”, describe five purposes for mixed-methods research,(77) which are defined as follows:
1. Triangulation: “seeks convergence, corroboration, correspondence of results from the different methods” (77)[p259]

2. Complementarity: “seeks elaboration, enhancement, illustration, clarification of the results from one method with the results from the other method” (77)[p259]

3. Development: “seeks to use the results from one method to help develop or inform the other method…” (77)[p259]

4. Initiation: “seeks the discovery of paradox and contradiction, new perspectives of frameworks, the recasting of questions or results from one method with questions or results from the other method” (77)[p259]

5. Expansion: “seeks to extend the breadth and range of inquiry by using different methods for different inquiry components” (77)[p259]

The quantitative component of this study was designed to examine the quantitative psychometric properties of the simulation assessments and MCQ tests, specifically: construct validity, formative validity, internal consistency and inter-observer reliability (inter-observer reliability applied only to the simulation assessments). The specific statistical approaches used to evaluate each of the psychometric properties are discussed in the relevant sections.

The qualitative component of this research project was employed for the purposes of complementarity and expansion. (77) In the case of formative validity, review of the literature did not identify any methods for evaluating this parameter, so the methods used to determine formative validity in this research are without precedent. It was determined that while statistical measures may give some indication of this formative validity, qualitative evaluation would provide complementary information, so both methods were used to evaluate this psychometric property. Conversely, it was decided
that qualitative inquiry into content validity would provide a more meaningful
evaluation than quantitative measures, and that a quantitative evaluation of content
validity would not be necessary. (50, 52, 78) Therefore, as per the concept of mixed-
methods for the purpose of expansion, a qualitative approach alone was used to assess
content validity. (77)

A pseudo-fieldwork approach was used to gather qualitative data during the research
process; the approach is termed pseudo-fieldwork because the method used was
modified from the traditional fieldwork approach. As per standard fieldwork practice,
qualitative observations were made by the researcher, and documented in field notes,
without “predetermined constraints on findings” (50)[p420]. However, the observations
were made in an environment/situation controlled and manipulated by the researcher:
participants were recruited and attended specifically to complete an MCQ assessment
and simulation assessment, and participants were allocated to one of two simulations by
the researcher. This is a departure from qualitative research using a true fieldwork
approach, which implies naturalistic inquiry, i.e. observation of occurrences in real
world situations, without researcher control or manipulation. (50)

As the field notes were acquired and reviewed, an abundance of data on the research
process and simulation assessment execution became apparent. In response to this
additional data, examination of the research process was added as a tertiary objective of
the study.
B. Research Process

Execution of this study involved four stages. Figure 3.1, below, shows each of these stages and an outline of the process followed for this research. Please note that, while data analysis is specified as the fourth and final stage, the qualitative component of data analysis was ongoing throughout the research process, from the onset of stage 1.

![Figure 3.1: Research process outline](image)

**Stage 1: Assessment tool development**

The first stage in the research process was assessment tool development. There are no standardised or validated MCQ tests or simulation assessment tools available in the area of cardiac emergency management, so these needed to be developed specifically for this research project. In order to develop high quality assessment tools, the MCQ and simulation assessments were designed, reviewed by experts, piloted and modified before they were finalized and validated. (79) The processes used to develop the assessment tools are described in more detail in *Chapter 4: Assessment Tool Development*. 

55
The second stage was participant recruitment. This involved advertising to potential participants, providing detailed information and obtaining informed consent. (50)

Participants were recruited using a purposeful sampling technique from the two population groups being studied—medical students and junior doctors. Medical students were recruited from Northern Clinical School (Northern Health, Victoria). This population included second, third and fourth year medical students (47 MD2, 44 MD3 and 44 MD4) from the University of Melbourne Doctor of Medicine course (a four year, graduate entry medicine course). Junior medical officers (JMOs) were recruited from the population of doctors in their first and second postgraduate years (40 intern and approximately 60 HMO2) employed by Northern Health. It is important to note that the actual available population from which to recruit was much smaller than the total population indicated above, as many students and JMOs were rotating to other sites during all or part of the recruitment period. It was not possible to determine the exact number of potential participants that were placed at Northern Health during this period, however, it was noted that only 10 of the 44 MD4 students were placed at Northern Health during this time, and just under half of the HMO2s are usually rotated to Northern Health at any one time.

A recruitment email was sent out to all medical students via the Northern Clinical School, and to all Northern Health JMOs via the Medical Workforce Unit. This email included information regarding the study, including aims, details of participation and estimated time commitment, and was approved in advance by the Northern Health HREC. The primary investigator did not have access to the contact details of potential participants unless the individuals themselves gave them.
Flyers with attached sign up lists were also placed in the Northern Clinical School, education precinct, resident quarters and doctors write up areas throughout The Northern Hospital to facilitate recruitment. Another method for recruitment was word of mouth, with students and junior doctor encouraging their peers and colleagues to become involved, either via face-to-face communication, or sometimes using methods of communication such as text messaging, Facebook, WhatsApp and the hospital paging system. Some senior staff members with an interest in the research project also encouraged medical students and junior doctors to become involved.

Participation in this study was completely voluntary and, while some senior staff and educators did promote the project to potential participants, there was no obligation, nor incentive, to participate. All participants were provided with a participant information sheet and were given a verbal explanation of what the study would involve, prior to signing a participant consent form. Participants were provided with a withdrawal of consent form, and informed that they could withdraw from the study at any time (Appendix 1: Participant Information Form and Participant Consent Form).

Medical students had all had some previous experience with simulation as a learning tool, as part of the teaching methods employed in the University of Melbourne MD course. They had not previously been formally assessed using simulation-based assessment, but had encountered OSCEs using standardised patients as part of their formal summative assessments.

Junior doctor participants had a range of previous experience with simulation, depending on the medical school and clinical school that they attended, and they may have had exposure to simulation through voluntary participation in other formal
courses (eg. Advanced Life Support). All junior doctors had been assessed for their Basic Life Support (BLS) competency (a condition of employment at Northern Health) using a brief low fidelity simulation-based assessment; this assessment usually involves some clinical context, but is sometimes simplified to demonstration of BLS skills only. It is assessed as pass/fail only and does not include a formal debrief. Nonetheless, all junior doctor participants had some familiarity with the use of simulation for assessment.

iii. Stage 3: Data Collection

Stage 3 involved data collection using multiple methods. Participants completed a participant questionnaire (to gather demographic data), an MCQ test, and a cardiac emergency simulation assessment (individually). Consideration was given to randomising participants to complete either the simulation or the MCQ test first, but this was not feasible due to small participant numbers; this is discussed further in Chapter 26: Future Directions. The results of the questionnaire, MCQ test and simulation assessment provided the quantitative data for this study. Qualitative data was collected through observation and collection of field notes.

Participant Questionnaire

Participants were required to complete a Participant Questionnaire to gather information on demographics, education and experience (Appendix 2: Participant Questionnaire). Demographic data included age, gender and year level (for medical students) or postgraduate year (for JMOs). Questions also addressed previous education in medical-related fields and previous clinical experience (experience in non-medical capacity; emergency medicine/ cardiology/ cardiothoracic surgery rotations). The questionnaire consisted of nineteen items in total and participants were directed as to which sections
they were required to complete, as some sections were only relevant to particular classifications. The participant questionnaire was not piloted prior to use in the research study; however, it was reviewed by education experts to evaluate the face and content validity.

Multiple Choice Question Test

Participants completed a multiple choice question (MCQ) test, consisting of thirty questions completed in a time limit of thirty minutes (Appendix 3: Multiple Choice Question Test). The questions were designed to assess knowledge on the topics of AMI and APO. Each participant completed the MCQ test prior to his or her participation in the simulation assessment. Information on the MCQ test development and content is included in Chapter 4: Assessment Tool Development.

Simulation Assessment Tool

Participants (n = 38) were allocated to complete either Simulation Scenario 1 (n = 19) or Simulation Scenario 2 (n = 18), which both used a live actor in a hybrid set-up. One participant completed the participant questionnaire and MCQ test but did not complete the simulation assessment. Figure 3.2, below, shows the division of participants into two groups for the simulation scenarios.

Allocation to simulation assessments was by simple alternate allocation to Scenario 1 or Scenario 2- i.e. participants were allocated participant identification numbers (PINs) according to the order in which they participated; participants with odd numbered PINs completed Scenario 1 and participants with even numbered PINs completed Scenario 2. This approach to allocation was used in order to simplify the process for actors and
The study was limited to two alternate scenarios due to resource availability and expected number of participants, in order to increase sample size for each scenario (compared to dividing participants across three scenarios). The two simulation scenarios depicted commonly encountered cardiac emergencies in the hospital setting, specifically AMI and APO. It was decided that of the three major cardiac emergencies (AMI, APO and cardiac arrest), that cardiac arrest should be excluded as cardiac arrest is comparatively simple to diagnose, has a very clear management algorithm which does not involve much subjective assessment, and is already commonly assessed through simulation, for example, in basic life support training.

The simulation scenarios, and their corresponding simulation assessment tools, were developed for this research project. Information on the simulation assessment development and content is included in Chapter 4: Assessment Tool Development. Detailed information regarding the simulation execution was provided to all assessors and actors. The simulation assessment tools were made available to the assessors prior to assessment of their first simulation, so that they could clarify any questions or
concerns with the student researcher. The checklist was marked out of a total of 27. The global assessment score (GAS) was marked from zero to three. Clear criteria were defined for each allocated mark.

Two assessors scored simulations in real-time, using scenario-specific assessment tools. Since inter-observer variability was specifically being examined in this research, each simulation score sheet was marked with an assessor identification number, so that results for each assessor could be examined (multiple assessors were required, since the simulations took place over several days). Importantly, one assessor was constant for all simulations, to serve as a baseline or control. The simulation assessment included a debriefing component to allow for reflection and specific feedback.

The possibility of recording the simulations was considered, which would have allowed multiple assessors (more than two) to assess each simulation remotely. A decision was made not to record the simulations, as there was concern that, in the setting of voluntary participation, the prospect of being recorded might dissuade potential participants from agreeing to participate. It was also considered that recording the simulations would be difficult given the very limited human resources- i.e. one of the assessors would have to mange recording, and given that both assessors were already involved in resetting the room and administrative tasks between simulations, this would be very difficult.

Field Notes
The approach taken for the qualitative component of this research project is described as pseudo-fieldwork because the situations under observation were created by the research process, rather than observations being made in a pre-existing naturalistic inquiry environment/situation.(50) This may be considered a purposeful sampling approach,
focused on the phenomena of interest - the research process and execution of the simulation assessments (50).

The researcher kept field notes during the course of the research process for later analysis and interpretation. These notes took a range of forms, and were collected using a range of methods, as described below, and were combined into a single electronic document:

1. Diary/timeline of events
2. Observations during the research process (commencing after ethics approval)
3. Observations during the simulation assessments
4. Discussions with participants as part of the simulation debriefing
5. Informal feedback from assessors during/after simulation days
6. Informal feedback from participants about the research project

All the data forms used for the qualitative component of the research were in some way created by the researcher: the field notes were based on the observations of the researcher or discussions between the researcher and other ‘players’ in the research (participants and assessors); the other data forms analysed (assessment tools, Multiple Choice Question Test Feedback document, debrief instructions for the simulation assessments) were designed and developed by the researcher. Ideally, the qualitative analysis component of this research would involve use of some ‘external’ qualitative data, for example data gathered directly from participants, assessors, or others not involved in the research (educational experts, etc.). This is a limitation of the methodology used in this research, and occurred as a result of the later addition of a significant qualitative component looking at the research project. At the stage when this component was added, it was not possible to change the ethics submissions to include...
additional data collection from participants or assessors, and so it was decided that a pseudo-fieldwork approach would be used.

The limitations associated with the lack of depth and breadth in the qualitative data is discussed in *Chapter 25: Limitations*, and potential avenues to address this are discussed in *Chapter 26: Future Directions*.

**iv. Stage 4: Data Analysis and Interpretation**

Stage 4 involved analysis of the data collected.

For the quantitative data, statistical analysis was performed using Statistical Package for the Social Sciences (SPSS, IBM Corporation). A p-value of 0.05 was pre-determined as the level accepted for statistical significance for all statistical tests. The statistical methods used to determine each measure of reliability/validity are described in more detail in the relevant section(s), but can be summarised as follows:

1. Assessment results:
   - descriptive statistics for demographic factors
   - chi-squared tests and t-tests for participant groups
   - Pearson and Spearman correlation for comparison between results on different assessments

2. Construct validity:
   - Cronbach’s alpha for internal consistency
   - descriptive statistics for contrasted groups comparison
   - Kruskal-Wallis test and Dunn-Bonferroni post-hoc methods for comparison of performance across all classifications
3. Formative validity:
   - descriptive statistics for each assessment item
   - Mann-Whitney U test and Kruskal-Wallis test for comparison of group performance on each item

4. Interobserver reliability:
   - percentage agreement
   - Spearman correlation
   - Cohen’s kappa

For the qualitative components, various different analysis techniques were required, dependant on the data forms being analysed, the objective of the analysis, and whether or not there was a mixed-methods (combined quantitative/qualitative) approach for that particular component. Further detail on the specific qualitative analysis methods used is provided in the relevant sections, but an overview of the techniques used is described below.

In the analysis of formative validity, a mixed-methods approach was used, the quantitative component of which is described above. No previous methods to determine formative validity had been described in the literature, and so the approach used in this research project was novel and without precedent. The data sources for this analysis were the *Multiple Choice Question Test Feedback* document (*Appendix 4*) and the debrief instructions for the simulation assessments. For the formative validity component, the qualitative analysis and presentation of results was descriptive only. A novel approach, based on concepts of content analysis and deductive coding was used. The two data sources (*Multiple Choice Question Test Feedback* document (*Appendix 4*) and the debrief instructions for the simulation assessments) were analysed for findings
related to the four previously described aspects of formative validity, which were used as conceptual codes (or domains), on which to base evaluations of formative validity. These four aspects were: form, type, amount and utility (of feedback).

The analysis of content validity was purely qualitative, and was undertaken both prospectively and retrospectively. Again, the qualitative analysis and presentation of results for this component was descriptive, and involved an integrative (deductive and inductive) approach to content analysis, designed specifically to determine whether the intended learning objectives were adequately addressed by the assessment tools (a far more defined and less open research question than is usually addressed by qualitative approaches). The approach used to perform the analysis was based on the method described in the text by Johnson & Christensen,(50) which involved a series of steps to check for content coverage (this is further described in Chapter 6).

The major qualitative component of this research was designed to address the tertiary objective of the study, which was “to examine the research process and to identify the strengths, weaknesses, limitations and issues related to conducting simulation assessment research”. This component, being larger and less constrained in its specific objectives (a more open research question) employed a more traditional approach to qualitative analysis. For this component, a grounded theory approach was used to develop taxonomies describing the “key conceptual domains and essential conceptual dimensions of the domains”.(80) The qualitative data analysis process was ongoing from the initial stages of assessment tool design, through to completion of the research project. The student researcher performed all coding, such that the application of codes was consistent throughout the field notes. Inductive coding was used, and applied to the field notes until the point of theoretical saturation. The codes and subcodes derived
from the qualitative analysis were then used to develop two separate taxonomies describing participant recruitment and logistics, and simulation assessment execution, to address the specified research objective. Written and diagrammatic representations of the qualitative results were produced.

The results obtained for both quantitative and qualitative components of the research were interpreted together in order to draw conclusions and develop recommendations, as presented in the discussion and summary sections of this thesis.

C. Research Ethics

Principles of educational research ethics were considered in the design and execution of this research project, to minimise any potential risk to participants. The research project was categorized as low risk, with possible risks identified as risk of psychological harm or devaluation of personal worth, related to the stress of participating in simulated emergencies and possible unfavourable simulated scenario outcomes. It was predetermined that participants’ actions in the simulations would not result in death of the simulated patient, to avoid potential psychological trauma. The major benefit for participants was the individual feedback and teaching that they received. Overall, it was judged that the benefits of participation outweighed the potential risks.

In regards to participant privacy, data management complied with The Australian Code for the Responsible Conduct of Research. Hard copy data is stored in a locked filing cabinet and electronic data is stored on a secure network, with all files password protected. Electronic data was only transmitted by email between secure networks (Northern Health Network and UWA Network). Original data was only be accessed by those directly involved in the research- student researcher and supervisors. Those
involved in review of statistical analysis only had access to de-identified data. All
records will be retained for a minimum of 7 years after publication or project
completion, whichever is later. When data is destroyed after the minimum retention
period, this will be done using secure destruction services for non-digital data, and
permanent erasure for digital data.

Ethics approval was obtained from the University of Western Australia Human Ethics
Office (RA/4/1/7400), in accordance with the requirements of the National Statement
on Ethical Conduct in Human Research and the policies and procedures of the
University of Western Australia. Ethics approval was also obtained from the Northern
Health Low-Risk Research Ethics Committee (LR 14.2015). Registration of External
Ethics Clearance was also completed through the University of Melbourne Medical
Education Human Ethics Advisory Group (Ethics ID 1544477).

All participants were informed of the purpose and potential risks of participating in the
research project and gave informed consent. Participants were advised that they could
withdraw consent (Appendix 1: Participant Information Form and Participant Consent
Form).
D. Summary

This chapter has provided an outline of the methodology employed for this research project, including a description of the research approach, process, ethics, participant recruitment and data collection. Chapter 4: Assessment Tool Development will describe the way in which the MCQ test and simulation assessments were developed for the purpose of this research project. A detailed description of the statistical and qualitative analysis methods follows in Chapters 5-10.
CHAPTER 4: ASSESSMENT TOOL DEVELOPMENT

This chapter describes the process involved in the development of the MCQ test and simulation assessments, including the approach and references used, content covered, and review process.

A. Multiple Choice Question Test Development

The MCQ test was developed specifically for this research project, according to assessment principles. Prior to designing the MCQ test and simulation scenarios, learning outcome categories, based on the ACF, were determined for each of the two topics. The learning objective categories were common to AMI and APO, as they relate to general medical assessment and management skills. Specific learning objectives were determined for use in the simulation assessment, but as many of these were practical rather than knowledge-based, they were not used for development of the MCQ test. The learning objective categories identified were:

A. Patient assessment
   1. History
   2. Examination
   3. Investigation
   4. Diagnosis
   5. Prioritisation

B. Patient management
   1. Management
   2. Patient progress
An equal number of questions were allocated to each of AMI and APO (fifteen for each). Clinical vignettes/stems and questions were constructed by the student researcher to address each of the pre-identified learning outcome areas from the ACF. While it was determined that each of the learning outcome areas should be covered by the MCQ test, it was not considered necessary that there be an equal number of items for each category. This decision was based on the fact that some categories are more practical/skills-based, and therefore better assessed by practical assessment methods, and others are more theoretical/knowledge-based, and therefore able to be assessed by a written assessment. (11, 22)

Three options were given for each MCQ item, as this was determined by a comprehensive meta-analysis by Rodriguez (2005) to be the optimal number of options. Reviews have shown that it is generally not possible to provide more than two plausible distractor options, such that examinees will narrow the options given to the correct answer and two plausible alternatives. By decreasing the number of options, it is also possible to administer more MCQ items in the same testing time, resulting in improved content coverage. (22)

Questions were written to test knowledge/application of knowledge at the level expected at the end of a medical degree, using the MCQs in the Australian Medical Council’s *Handbook of Multiple Choice Questions* as a guide for the expected standard. (81) Similarly, some clinical experience, at the level of a graduating doctor, was assumed. That is, participants were assumed to be familiar with patient charts (e.g. medication charts), medical escalation codes (e.g. MET calls), pathology requests, radiology requests, etc.
While the targeted level was the end of final year medical school, the fact that more junior medical students, and more senior JMOs would be completing the test was taken into account. In particular, some questions of lesser difficulty were included so that the performance of the least experienced participants could still be differentiated, and so that all participants felt able to answer some questions. Similarly, some more difficult questions were included to take into account that some HMO2 participants had completed/were completing a specialty rotation in cardiology, and to allow differentiation between the highest performing participants. Inclusion of appropriate questions for the different levels of experience was evaluated during the process of MCQ review.

Sound, evidence-based resources were used in construction of the MCQ items and confirmation of correct answers. Where possible, local guidelines were used in the construction of items, to ensure that answers reflected best practice in the Australian context. Specific references used include:

- The National Heart Foundation of Australia/The Cardiac Society of Australia and New Zealand- Guidelines for the management of acute coronary syndromes 2006: Summary of key recommendations
- The Royal Melbourne Hospital Evidence Based Guidelines: Management of Acute Pulmonary Oedema
- Medscape articles on the relevant pathologies
  (Medscape is an online searchable resource for physicians and healthcare professionals- http://www.medscape.com)
After initial drafting and re-drafting of the test items, development of the test included multiple rounds of consultation with, and feedback from, content and medical education experts. These experts included two cardiologists, two medical education academics, and two medical educators. The aim of this review process was to ensure that there was adequate content validity and clarity of items. Each expert reviewed the MCQs in turn and offered suggestions; these suggestions included exclusion of certain items, inclusion of additional items (to cover particular topics/areas), alternative wording, simplification of item stems and responses, addition of further information to stems (e.g. vital signs), and comment on the level of difficulty (to accommodate the most junior and most senior participants). Overall, six experts reviewed and provided comments on the MCQ assessment. Modifications were made between reviews by different experts, including the addition, deletion and modification of items, until the set of MCQs was deemed suitable by the student researcher and both research supervisors.

The MCQ test was then piloted with a HMO2 doctor from another health service who was not eligible to participate in the study. A second external HMO2 doctor undertook a review of the assessment, in the same way as the experts reviewed the items, with answer keys available. Both of these colleagues provided feedback and suggestions for improvement to the MCQ test. Changes were subsequently made, to improve the clarity of the items, and to reduce reading material. Full-scale piloting of the MCQ test, which is recommended to improve validity, was not possible due to a limited pool of potential participants (as piloting the test on any students/junior doctors who were potentially eligible to participate would decrease the number of study participants) and time/resource limitations.
A feedback document for the MCQ test, with answers and comprehensive explanations, was also developed and was distributed to participants following completion of the data collection period via email, approximately two weeks after the last participants had completed their assessments (*Appendix 4: Multiple Choice Question Test Feedback*). Participants who did not provide their email address to the researcher, and who could not be emailed via the Northern Health internal email system, were forwarded the answers document via the Northern Clinical School.

**B. Simulation Assessment Development**

The aim of the simulation scenarios was to allow participants to gain experience in managing a common cardiac emergency situation in a safe simulated environment, with assessment undertaken to allow for individual feedback to be provided.

The researcher, in consultation with content and medical education experts, developed the scenarios used for the simulation. As for the MCQ test, content for these assessments were drawn from relevant sections of the ACF (Australian Curriculum Framework for Junior Doctors). (82) The same learning outcomes categories were used in development of the MCQ test and the simulation scenarios. However, specific learning outcomes were identified for the simulation assessments, again, based on the ACF. The specific learning objectives are listed under the simulation scenarios described below.

The same, evidence-based resources that were used in construction of the MCQ test were also used in development of the simulation scenarios and simulation assessment tool. Where possible, local guidelines were used in the construction of items, to ensure
that answers reflected best practice in the Australian context. The assessment sheets were structured to reflect the learning outcomes and listed under the same headings as the MCQ learning outcome categories, though prioritisation was included in patient management, rather than patient assessment. The checklist was marked out of a total of 27. The global assessment score (GAS) was marked from zero to three. Clear criteria were defined for each allocated mark. The maximum achievable score for the simulation assessment was 30, the same as for the MCQ test.

Given the resources available, the simulations were designed with live actors (patient and nurse) in a hybrid set-up, with each scenario running for a maximum of ten minutes plus five minutes debrief (participants could end the scenario early if they felt that there was nothing else they would like to do). Participants were instructed to assume the role of a junior doctor in managing the emergency.

The simulation scenarios document included detailed information for the actors on how to respond to questions and how to portray their clinical condition (e.g. portraying tachypnoea). Unfortunately, different actors were used during the research project, as it was not possible to secure the same actor for all simulations sessions. Very simple moulage was used to simulate cellulitis for Scenario 2 (make up used to simulate erythema). Actors playing the part of nurses were trained nursing or medical staff, and were advised to imitate performing nursing tasks in real time (e.g. it would take approximately one minute to insert a cannula or obtain an ECG). They were advised not to prompt participants except for specific questions to assist in assessment (e.g. asking what the ECG showed, whether or not the participant would like blood tests to be taken at the time of intravenous cannula insertion).
The way in which vital signs and patient symptoms would change with institution of different management strategies was pre-determined and included in the simulation scenarios document (Appendix 5: Simulation Scenarios). The sequence of events throughout each simulation was dependent on the actions taken by participants. Patient vital signs started at set levels and were shown dynamically on a “monitor”, (a remotely controlled iPad) next to the patient’s bed. Participants were advised to examine the simulated patient and that the examiner would provide examination findings. If investigations were ordered, the results became available in real time (i.e. an ECG or blood gas would become available quickly, while a formal pathology result or chest x-ray would not be available during the simulation). It was pre-determined that, regardless of the actions of the participants, death would not be an outcome of either simulation. Each simulation experience was strictly ten minutes.

The simulation scenarios document included instructions for assessors to provide a structured debrief, which was an opportunity for an interactive feedback discussion with each participant. Each debrief lasted approximately five minutes. Development of the simulations included multiple rounds of consultation/review with content and medical education experts, in order to ensure that there was adequate detail, content validity and that the scenarios were accurate representations of the medical conditions (AMI and APO).(50, 59, 76, 78) The review process used was the same as for the MCQ test, but included review by four, rather than six, experts.

A test run of each of the simulation scenarios, without props or simulated nurse, was undertaken by a volunteer HMO2 doctor from another health service who was not eligible to participate in the study. Subsequent discussion resulted in some very minor changes to minimise any ambiguity in the simulated scenarios. In particular, more
emphasis was put on portrayal of the clinical situations/pathologies through the objective features of the simulations (vital signs, ECGs), since subtle patient cues are difficult to portray through simulated patients.

i. Simulation Scenario 1

Simulation Scenario 1 depicted a 62-year-old patient (flexibility in gender of the patient was required to allow for different actors to be used as simulated patients) presenting to the emergency department with chest pain. The scenario was designed as a typical presentation of AMI, with the patient giving a description of typical ischaemic chest pain and associated symptoms and appropriate risk factors for ischaemic heart disease, such that participants should have recognised a high pre-test probability of myocardial ischaemia/infarction prior to any investigations. The ECG, if requested, demonstrated obvious anterior ST elevation meeting STEMI criteria, with reciprocal changes particularly obvious in inferior leads. Chest pain improved with appropriate treatment.
The specific learning outcomes for Simulation Scenario 1, based on the ACF, were:

**A. Patient assessment:**
1. History
   - takes a focused history to differentiate between causes of chest pain
2. Examination
   - performs a focused cardiorespiratory examination
3. Investigation
   - requests appropriate investigations
   - interprets an ECG
4. Prioritisation
   - recognises patients who are acutely ill and require immediate treatment
   - recognises when to call for help
5. Diagnosis
   - is able to form a list of differential diagnoses and determine the most likely diagnosis

**B. Patient management:**
1. Management
   - identifies and initiates appropriate initial management for acute coronary syndrome
   - identifies the need for senior clinician input to determine and implement further management
2. Patient progress
   - re-evaluates the patient’s progress and response to treatment

**ii. Simulation Scenario 2**

Simulation Scenario 2 depicted a recently admitted 76-year-old patient with cellulitis, with a new complaint of dyspnoea, for whom a pre-MET call was activated due to symptomatic tachypnoea and desaturation. The scenario was designed to portray a common cardiac emergency situation, where a series of events has led to fluid overload and APO. In this scenario, the patient’s admission was for cellulitis and, due to the infection, pre-existing AF was unmasked/worsened. As a result of the AF, and intravenous fluids given, the patient developed APO. The history given by the patient, on questioning, included dyspnoea (non-specific) and palpitations (suggesting AF). Importantly, the past medical history included in the patient folder included AF. Examination findings gave an obvious impression of significant fluid overload. The ECG, if requested, demonstrated obvious AF with rapid ventricular rate. Dyspnoea
improved slightly with oxygen, morphine or intravenous frusemide but not with oral frusemide.

The specific learning outcomes for Simulation Scenario 2, based on the ACF, were:

**A. Patient assessment:**
1. History
   - takes a focused history to differentiate between causes of dyspnoea
2. Examination
   - performs a focused cardiorespiratory examination
3. Investigation
   - requests appropriate investigations
   - interprets an ECG
4. Prioritisation
   - recognises patients who are acutely ill and require immediate treatment
   - recognises when to call for help
5. Diagnosis
   - is able to form a list of differential diagnoses and determine the most likely diagnosis

**B. Patient management:**
1. Management
   - identifies and initiates appropriate initial management for acute pulmonary oedema/fluid overload
   - identifies the need for senior clinician input to determine and implement further management
2. Patient progress
   - re-evaluates the patient’s progress and response to treatment

**C. Summary**

This chapter has described the way in which the MCQ test and simulation scenarios were developed, including the design and review processes. The next six chapters of this thesis (*Chapters 5-10*) address the methods used for evaluation of participant characteristics and overall results, the different psychometric properties, and the research process.
CHAPTER 5: ANALYSIS OF PARTICIPANT CHARACTERISTICS AND ASSESSMENT RESULTS

Descriptive statistics were calculated for participant demographics, including gender, age, classification, course/university, undergraduate/graduate degree. Chi-squared tests and independent samples t-tests were used to assess for any significant differences between the participants allocated to Scenario 1 and Scenario 2 groups. Additional demographic data collected through the participant questionnaire, including data on previous education and experience, was initially intended to allow subgroup analysis. However, low participant numbers did not allow for these subgroup analyses to be conducted.

Descriptive statistics were also calculated for the MCQ test and simulation assessments. Participant scores were represented graphically on a histogram to assist in interpretation of the distribution. The Shapiro-Wilk test was used to numerically analyse the distribution of results, since this is the most powerful normality test, and appropriate for the sample size.

The correlation between the checklist and GAS (global assessment score) components of the simulation assessment tools was also assessed. Since the GAS results were not normally distributed, correlation was calculated using the Spearman correlation coefficient ($\rho$), a non-parametric test for correlation. The purpose of this part of the analysis was to determine whether a very defined checklist score, marked strictly based on execution/non-execution of critical tasks, would correlate with a general assessment score, designed to assess overall performance.
In analysing the results of the simulation assessments, the results obtained by Assessor 1 were used. Assessor 1 marked the performances of all participants, whereas the second assessor was variable, and used purely for the purpose of determining interobserver reliability. Had the second assessor also been constant, the mean marks of both assessors could have been used in the analysis of participant results. However, use of a mean mark was considered to be inappropriate in the setting of multiple different second-assessors, where the effect of particularly stringent or lenient second-markers could have affected different participants’ scores variably (a large inter-marker effect). (86) The use of a single examiner’s marks is also prone to issues, in particular, marker ‘drift’ (trend towards more stringent or lenient marking over time) or instability (both intra-marker effects). (86) However, it was considered that, overall, the cumulative effect of one marker’s intra-marker effects is likely to be less than the combined effects of multiple inter-marker effects AND intra-marker effects.

Pearson and Spearman correlations were used to determine the correlation between performances on the MCQ test and the simulation scenarios. Analyses were undertaken for each simulation scenario, and for the components of the simulation assessments (checklist component, GAS), to determine if these were better correlated with performance on the MCQ test. Pearson correlations (r) were used for comparisons of total score and checklist score, since these scores were shown to be normally distributed; Spearman correlations (ρ) were used for comparisons that involved the GAS component, since this data was not normally distributed. (87) It was decided that the simulations scenarios should be considered separately, rather than combined, since the distribution of scores was different for the two simulation scenarios.
The analysis of correlation between results of the MCQ and simulation assessments was designed to address the research question that inquires as to the relationship between performance on simulation-based and MCQ assessments of cardiac emergency management (Research Question 2)- “What is the relationship between performance on simulation-based and MCQ assessments of cardiac emergency management? What is the correlation between participants’ scores on a simulation-based assessment and an MCQ test both designed to test cardiac emergency management?”.
CHAPTER 6: CONTENT VALIDITY

The research methods described in this chapter were designed to address Research Question 1E—“What is the content validity of each method of assessment?: To what extent does each method of assessment evaluate achievement of the relevant learning objectives, as documented in the ACF?” (51) Both quantitative and qualitative methods to evaluate content validity have been developed and used in psychometric analyses. However, there appears to be no clear consensus as to which methods are most accurate. (52) Quantitative methods are resource-intensive and require formal review and scoring of a large pool of potential test items by multiple content experts. Measures such as the content validity ratio (CVR) and content validity index (CVI) can be calculated and, on this basis, items included in, or excluded from, an assessment tool based on these measures. (52)

Quantitative analysis is not specifically necessary to evaluate content validity, and adequate analysis may be achieved by qualitative methods. (50, 52, 78) It is appropriate that content, being a conceptual rather than numeric property, is assessed qualitatively. As such, qualitative methods have been employed in this study to assess the content validity of both the MCQ test and simulation assessment tool. The analysis of content validity was undertaken both prospectively, during the assessment tool development phase, and retrospectively, after the conclusion of the data collection phase, as described below.

The qualitative analysis of content validity was based on a content analysis approach, following the methods described by Johnson and Christensen in the textbook Educational Research: Quantitative, Qualitative, and Mixed Approaches, and was purely descriptive. (50) However, the method described in the text was rather general
and not specific in the details on methods for analysis. As such, in order to add methodological rigor to this approach, in integrative approach to developing code structure was used: first deductive coding was to ensure that all learning objectives were covered, followed by inductive coding to ensure that no items were irrelevant.

Since content validity requires that there is adequate sampling of the aspects being tested, it was necessary to first clearly define what was being assessed. It was decided, prospectively, that the MCQ test and the simulation scenarios would assess two specific topics in cardiac emergency management: AMI and APO. The MCQ test was to test both of these topics in one assessment tool, whereas each of the two simulation scenarios assessed one of these topics.

The first step in analysis of content validity was to determine whether the items appeared to represent the content being assessed.(50) This step was addressed in the initial assessment tool development and stages of expert and peer review and any changes required were made to ensure that this criterion of content validity was fulfilled. The second step was to determine the relationship of the test items to the learning objectives. This involved deductive coding, with each of the MCQ test items coded by learning outcome category and topic (AMI or APO), to evaluate whether or not, and to what degree, each learning objective was sampled by the test.(50)

For the simulation assessment, a similar deductive coding approach was used, however each the items were coded by specific learning outcomes rather than just learning outcome categories, since the simulation scenarios were actually designed based on specific learning outcomes. If the assessment tools had been designed to assess achievement of learning outcomes taught in a particular course of instruction, then it
would be pertinent to also evaluate alignment between proportion of instructional time to proportion of items assessing that content at this point. However, this was not relevant in this context.

The third step in analysis for content validity was to determine if there were any irrelevant items included in the MCQ test. While this could be achieved by simply checking over items to ensure that all were relevant, it was decided that a more comprehensive approach should be used. All items were allocated keywords (usually multiple) for coding purposes, using an inductive approach, then checked to ensure that either AMI or APO, or an adequately related term had been identified as a keyword (code) for each item. The keywords that were determined to show adequate relevance were: AMI, APO, ACS (acute coronary syndrome), CCF (congestive cardiac failure), chest pain, HF (heart failure), IHD (ischaemic heart disease) and dyspnoea.

This third step was not applied to the simulation assessments because in clinical practice, it is necessary to consider the situation broadly and identify and exclude other differential diagnoses. Therefore, some items that are included in the simulation assessment tool will not be specific to the topic (AMI or APO) and may be part of a more general cardiorespiratory assessment/workup; as such it is not possible to prove their relevance using keywords.
CHAPTER 7: CONSTRUCT VALIDITY

The statistical methods described in this chapter were designed to address Research Question 1A- “What is the construct validity?: To what extent does each method of assessment actually measure the construct of interest, that is, participants’ competence in managing cardiac emergencies?(51)

A. Internal Consistency

Internal consistency reliability refers to correlation between test items within an assessment.(50-52, 88, 89) Internal consistency is a measure of the consistency with which items measure a construct, so the aim in evaluating the internal consistency is to determine whether the items in an assessment are related, i.e., whether they are assessing the same construct.(50, 88) For tests that assess multiple constructs, low internal consistency is expected, since the assessment is heterogenous. In this case, the assessment can be considered as multiple components, and the internal consistency should be determined for each component. Homogeneity, and therefore high internal consistency, would then be expected within each component.(50-52)

Cronbach’s alpha, also referred to as coefficient alpha, is the standard measurement used to determine the internal consistency of an assessment tool. (23, 50, 52, 78, 88)

Cronbach’s alpha is usually calculated using the following formula:(89)

\[ \alpha = \left( \frac{k}{k-1} \right) \left( 1 - \frac{\Sigma \sigma_i^2}{\sigma_y^2} \right) \]

Where:  
\( \alpha \) is coefficient alpha 
\( y \) is a multi-item scale 
\( k \) is the number of items in the scale, \( y \) 
\( \Sigma \sigma_i^2 \) is the sum of the variances for the individual items 
\( \sigma_y^2 \) is the variance of the scale as a whole
Cronbach’s alpha gives a result between 0 to 1, where a higher value implies greater homogeneity among the items, and lower index of measurement error. (50, 88) However, Cronbach’s alpha is also affected by the number of items, such that an increase in the number of items can increase the coefficient alpha without necessarily improving homogeneity, and a test with fewer items may have a lower coefficient alpha despite high homogeneity among the items. (50, 52, 88, 89)

A simplified version of Cronbach’s alpha, which demonstrates the concept, and the effect of the number of items, is shown below: (50)

\[
\alpha = \frac{k\bar{r}}{1 + k\bar{r} - \bar{r}}
\]

Where:
- \(\alpha\) is coefficient alpha
- \(k\) is the number of items
- \(\bar{r}\) is the average correlation between the items

Cronbach’s alpha was determined for the entire MCQ test and each simulation scenario, as well as separately for the AMI items and the APO items on the MCQ test, as these could be considered as separate components of the assessment. (52) Some items were excluded from analysis of Cronbach’s alpha, as there was zero variance. These were: MCQ item 20, Scenario 1 items A1e, A5a, A6a and Scenario 2 items A1b, A3d.

B. Contrasted-Groups Approach

The contrasted-groups approach was used as another method of determining construct validity. (52) This method involves comparing the performance of the most junior and most senior participants, and is based on the concept that experts would be expected to achieve significantly superior results to novices if the correct construct (i.e. ability to manage cardiac emergencies) is being assessed. (52, 73, 75, 76)
Descriptive statistics were calculated to determine the MCQ test and simulation assessment result profiles for each of the groups, based on classifications (MD2 through to HMO2).(50, 83) Given the small number of participants and more uniform performance within each comparison group, non-parametric statistical analysis was required to determine the significance of any differences in performance between groups.(83)

The MD2 and HMO2 groups were used to represent ‘novice’ and ‘expert’ groups, respectively. The Mann-Whitney U test, which is a non-parametric test similar to the independent-samples T-test, was used to compare performance rankings for these two comparison groups.(83) The two-tailed asymptotic significance is used as a measure of statistical significance of the difference between groups compared using the Mann-Whitney U test. If the asymptotic significance is less than the pre-determined significance level (0.05), this implies a statistically significant difference in performance between the two comparison groups (MD2 and HMO2).(83)

C. Comparison Across All Groups

A previous study by Lypson et al used the contrasted-groups method to validate an assessment tool for aseptic technique in resident physicians and showed a significant difference in performance between novice and experts, but no difference between the PGY1 group and a mixed PGY2/3 group.(75) This raised some concern that differential results between novices and experts does not necessarily correlate to the ability of an assessment tool to distinguish between examinees of a more similar level. In order to address this concern, Lypson et al also performed a comparison of performance across the full range of participant classifications, in order to determine whether there were
distinguishable differences in performances with less pronounced differences in experience. (75)

The Kruskal-Wallis test and Dunn-Bonferroni post-hoc methods were to compare performance across all classifications (MD2, MD3, MD4, intern, HMO2). The Kruskal-Wallis test was used to determine if there is an overall difference in performance across classifications and then the Dunn-Bonferroni post hoc method was used to establish differences more specifically between each different combination of groups. (90) The Kruskal-Wallis test is the non-parametric equivalent to the ANOVA (analysis of variance) test, and tests the null hypothesis that the distributions within the groups are the same. If the calculated two-tailed asymptotic significance is less than the pre-determined significance level (0.05), then the null hypothesis is rejected, implying that there is a significant difference between the groups. (83)
CHAPTER 8: FORMATIVE VALIDITY

The methods described in this chapter were designed to address Research Question 1B—“What is the formative validity?: To what extent does each method of assessment provide information that can be used to inform further participant training/education in management of cardiac emergencies?” (51)

A. Methods for Evaluating Formative Validity

Formative validity is an area of assessment evaluation that is poorly understood and inadequately described in the education literature. To the extent able to be determined by literature review, no methods of determining the formative validity of an assessment tool have been described. As such, the methods used to evaluate formative validity in this study are without precedent, and are simply methods that have been identified to potentially gather the information that is desirable from a formative assessment.

It was decided that a mixed quantitative and qualitative approach would provide the greatest amount of information by which to evaluate the formative validity of the assessment tools. The quantitative analysis for formative validity focused on identifying items on the assessment items that could be used to differentiate between candidates and therefore guide future learning. Qualitative analysis focused on evaluating the feedback that was provided to participants though the assessments (MCQ and simulation assessments).
B. Quantitative Component

The mean score and standard deviation were calculated for each item on each assessment tool, with mean score (mean performance) providing a measure of item difficulty. The spread of items by mean performance was then determined. Items were also sorted by content and difficulty, to evaluate the way in which different content areas were assessed.

Overall participant performance and group (classification) performance on individual items on the assessments were analysed using the Mann-Whitney U and Kruskal-Wallis tests, including Dunn-Bonferroni post-hoc correction for pairwise comparisons using the Kruskal-Wallis test. These tests were used in a similar way to their use for evaluating construct validity: the Mann-Whitney U test was used to determine which items could distinguish between ‘novices’ and ‘experts’. The Kruskal-Wallis test with Dunn-Bonferroni correction was used to identify items with the ability to distinguish between participants of more similar levels. These statistical tests and their purposes were described in Chapter 7: Construct Validity.

The statistical methods described above were chosen because they allow determination of the areas in which examinees performed particularly well or poorly, and identification of items with an ability to distinguish between candidates of different levels. This could provide very useful information to educators if these assessments were to be used in the context of an educational program.(16, 51)
C. Qualitative Component

A qualitative analysis was undertaken to evaluate the feedback provided through the MCQ test and simulation assessments, in order to determine and compare the formative validity for each assessment method. The data sources for this analysis were the *Multiple Choice Question Test Feedback* document and the debrief instructions for the simulation assessments. Qualitative analysis of all of these data sources was undertaken, using a novel approach based on conventional qualitative analysis methods, specifically content analysis and deductive coding.

Since the source of the qualitative data for this section was *not* the field notes, but rather other documents that formed part of the research materials, a true coding approach to analysis of all the data was not feasible. Instead, using an approach based on deductive coding, the content of the qualitative data sources was analysed, using specific headings as domains (similar to codes). Literature relating to formative validity has identified four key features of feedback which determine the formative validity of a given assessment method, (91, 92) and these were used as the domains:

- Form of feedback
- Type of feedback
- Amount (quantity) of feedback
- Utility (quality) of feedback.

This evaluation was intended to be descriptive, to allow for comparison between the two assessment methods and evaluation of overall formative validity, rather than interpretive.
CHAPTER 9: INTEROBSERVER RELIABILITY

The statistical methods described in this chapter were designed to evaluate Research Question 1D- “What is the inter-observer reliability of cardiac emergency simulation as an assessment tool?: To what extent does scoring on a simulation-based assessment of cardiac emergency management vary between assessors?”(51)

Three measures of interobserver reliability were calculated- percentage agreement, Spearman correlation and Cohen’s kappa.(23, 50, 72, 74, 75, 93) The measure of percentage agreement (i.e. the proportion of cases in which the examiners awarded the same score to the examinee) has the benefits of being easily calculable and easily understood conceptually, though not necessarily easily interpreted. However, this measure alone is inadequate to describe the interobserver reliability, therefore other measures were also employed.(94-96)

Spearman correlation was used because correlation coefficients are able to serve as a relatively simple measure of the relationship between two assessors’ scores.(50) Spearman correlation was used, rather than Pearson correlation, because the data for each combination of assessors is based on very small numbers, and the non-normality of this data would unduly influence the Pearson correlation, but not the Spearman correlation which uses ranks of scores rather than raw scores.(87) Cohen’s kappa was also used because it “takes into account chance agreement between two or more observers”, and “has become one of the preferred statistical methods for calculating the degree and significance of agreement between observers”.(95)[p209] For each of these methods, each combination of assessors was examined separately, since none of the above measures of interrater reliability is able to account for different combinations of assessors for different cases.
For the Spearman correlation, calculations were performed for each simulation scenario separately, but not for both simulation scenarios combined. This is because the Spearman correlation uses ranks, and, due to differences in distribution/range of results for each scenario, the results of both scenarios cannot be ranked together without first standardising results (which would potentially then result in compounded error from two different statistical methods).(87) For percentage agreement and Cohen’s kappa, measures of interobserver reliability were obtained for each simulation scenario separately and for both simulation scenarios combined. This was possible because the agreement between assessors would not be expected to be significantly different for two assessment tools with the same format.

Assessor 4 was not included in the analysis, since this observer only assessed two participants, and therefore there is inadequate data for statistical analysis. Similarly, interobserver reliability was not calculated for the combination of Assessor 1 and Assessor 2 for individual scenarios, due to small numbers (three cases for Scenario 1 and two cases for Scenario 2). For three participants, only one assessor marked the simulation, and therefore these results are also excluded from calculations of interobserver reliability.

Percentage agreement between examiners was also calculated for each individual item on each simulation assessment, to determine whether there were specific items on which assessors were more likely to differ in their marking. Correlation coefficients and kappa values were not used, as these measures are designed to compare continuous variables, rather than binary scoring (i.e. checklist items, which are assessed as completed/not completed).(50, 58) The percentage agreement for each item was calculated for each combination of assessors, as well as overall (for any combination of assessors).
CHAPTER 10: QUALITATIVE ANALYSIS OF THE RESEARCH PROCESS

The qualitative methods described in this chapter were designed to address Research Question 4- “What are the strengths, weaknesses, limitations and issues related to conducting simulation assessment research?”

The field notes that were recorded during the research project related to two separate topics: participant recruitment and logistics, and simulation assessment execution. These two separate topics were identified early in the process, and so field notes were recorded separately (under separate sections of the electronic field note document) for these two topics. This avoided the need to re-classify the observations made into the two topic areas during qualitative analysis.

Qualitative analysis was performed for each of these two topics with the intention of generating taxonomies, defined as “a formal system for classifying multifaceted, complex phenomena … according to a set of common conceptual domains and dimensions.”(80)[p1760] The qualitative data analysis process was ongoing from the initial stages of assessment tool design, through to completion of the research project.

For the topic of participant recruitment and logistics, the focus was to identify:

1. Factors that facilitated, limited, or influenced participant recruitment
2. Practical considerations in the planning, recruitment and execution phases of the research project

On the topic of simulation assessment execution, there was a broader focus, which was to group together the many observations made in regards to what occurred during the simulation assessments.
The first step in analysis was reviewing the raw filed notes to develop an understanding of the type, quality, quantity, breadth and depth of the data that had been collected. (80)

Coding of the data was then performed using an inductive/grounded theory approach “by which themes and categories emerge from the data through the researcher’s careful examination and constant comparison.” (97)[p2] The student researcher conducted all of the coding, ensuring continuity and consistency in the assignment of the codes. (80)

Codes were applied to the data until the point of theoretical saturation.

The codes and subcodes derived from the qualitative analysis were then used to develop two separate taxonomies describing participant recruitment and logistics, and simulation assessment execution, which were the two phenomena specified by the research objective. Essentially, the domains identified paralleled the codes that had been applied, and the dimensions paralleled the subcodes. The domains and dimensions that were identified during analysis were then further analysed to identify any relationships or linking variables between them.

Written and diagrammatic representations of the qualitative results are presented in Chapter 16, and expanded upon in Chapter 22.
SECTION III: RESULTS

CHAPTER 11: PARTICIPANT CHARACTERISTICS AND ASSESSMENT RESULTS

A. Participant Characteristics

Thirty-eight participants were recruited for this research project. Of these, 17 were students and 21 were JMOs. All 38 participants completed the multiple choice test and 37 of the participants completed the simulation assessment; 19 participants were allocated to Scenario 1 and 18 participants were allocated to Scenario 2. Sixty-three per cent of participants (n = 24) were female. The average age of participants was 25.8 years. Of the eleven HMO2 participants recruited, there were seven from the medical stream, two general, one surgical and one other. Table 11.1 shows the demographic data of the participants overall, and for the groups allocated to Scenario 1 and Scenario 2. There was no statistically significant difference between the participant demographics of the two groups, though it was observed that a larger proportion of more junior participants (especially MD2 level) were allocated to Scenario 2 than to Scenario 1.

Almost three quarters of participants (n = 28, 73.7%) had completed or were completing the University of Melbourne MD course. Of those participants who were junior doctors, 52.4% (n = 11) most had gained their medical qualification from University of Melbourne, and 76.2% (n = 16) had completed a graduate entry medical degree. One participant had completed a medical degree at an interstate university and did not specify whether the course was undergraduate or graduate entry. Table 11.2 shows information on participant medical training.
Table 11.1: Demographic data of participants overall, and of the groups allocated to Scenario 1 and Scenario 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>All participants (n = 38)</th>
<th>Scenario 1 (n = 19)</th>
<th>Scenario 2 (n = 18)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>24</td>
<td>63.2</td>
<td>14</td>
<td>73.7</td>
</tr>
<tr>
<td>Male</td>
<td>14</td>
<td>36.8</td>
<td>5</td>
<td>26.3</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>25.8</td>
<td>2.9</td>
<td>26.3</td>
<td>2.6</td>
</tr>
<tr>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD2</td>
<td>8</td>
<td>21.1</td>
<td>1</td>
<td>5.3</td>
</tr>
<tr>
<td>MD3</td>
<td>6</td>
<td>15.8</td>
<td>4</td>
<td>21.1</td>
</tr>
<tr>
<td>MD4</td>
<td>3</td>
<td>7.9</td>
<td>3</td>
<td>15.8</td>
</tr>
<tr>
<td>Intern</td>
<td>10</td>
<td>26.3</td>
<td>6</td>
<td>31.6</td>
</tr>
<tr>
<td>HMO2</td>
<td>11</td>
<td>28.9</td>
<td>5</td>
<td>26.3</td>
</tr>
</tbody>
</table>

Table 11.2: Information on participant medical training

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>Primary Grouping</th>
<th>N</th>
<th>%</th>
<th>Secondary Grouping</th>
<th>N (%)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Participants: Medical Degree</td>
<td>UoM- MD</td>
<td>24</td>
<td>63.2</td>
<td>UoM</td>
<td>28</td>
<td>73.7</td>
</tr>
<tr>
<td></td>
<td>UoM- MBBS- Undergraduate</td>
<td>3</td>
<td>7.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UoM- MBBS- Graduate entry</td>
<td>1</td>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monash- Undergraduate</td>
<td>1</td>
<td>2.6</td>
<td>Monash</td>
<td>3</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td>Monash- Graduate entry</td>
<td>2</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deakin</td>
<td>6</td>
<td>15.8</td>
<td>Deakin</td>
<td>6</td>
<td>15.8</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1</td>
<td>2.6</td>
<td>Other</td>
<td>1</td>
<td>2.6</td>
</tr>
<tr>
<td>All Participants: Undergraduate/ Graduate Entry</td>
<td>Undergraduate</td>
<td>4</td>
<td>10.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graduate</td>
<td>33</td>
<td>86.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unspecified</td>
<td>1</td>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JMOs: Medical Degree (n = 21)</td>
<td>UoM- MD</td>
<td>7</td>
<td>33.3</td>
<td>UoM</td>
<td>11</td>
<td>52.4</td>
</tr>
<tr>
<td></td>
<td>UoM- MBBS- Undergraduate</td>
<td>3</td>
<td>14.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UoM- MBBS- Graduate entry</td>
<td>1</td>
<td>4.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monash- Undergraduate</td>
<td>1</td>
<td>4.8</td>
<td>Monash</td>
<td>3</td>
<td>14.3</td>
</tr>
<tr>
<td></td>
<td>Monash- Graduate entry</td>
<td>2</td>
<td>9.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deakin</td>
<td>6</td>
<td>28.6</td>
<td>Deakin</td>
<td>6</td>
<td>28.6</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1</td>
<td>4.8</td>
<td>Other</td>
<td>1</td>
<td>4.8</td>
</tr>
<tr>
<td>JMOs: Undergraduate/Graduate Entry (n=21)</td>
<td>Undergraduate</td>
<td>4</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graduate</td>
<td>16</td>
<td>76.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unspecified</td>
<td>1</td>
<td>4.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
B. Multiple Choice Question Test

Figure 11.1, below, demonstrates the distribution of MCQ scores. The mean score for the MCQ test was 23.0 (range 16-30), with a standard deviation of 3.4. The Shapiro-Wilk significance was 0.26, indicating approximately normal distribution of the results.

![Figure 11.1: Distribution of scores for the MCQ test](image-url)
C. Simulation Assessments

Table 11.3, below, demonstrates descriptive statistics for the simulation assessments—
for each component of the simulation assessment tools, and overall. The mean score for
Simulation Scenario 1 was 22.9, more than six marks higher than the mean score of
16.1 for Scenario 2.

<table>
<thead>
<tr>
<th>Simulation Scenario</th>
<th>Possible Score</th>
<th>Mean</th>
<th>SD</th>
<th>Range (Min-Max)</th>
<th>Shapiro-Wilk Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1</td>
<td>0-30</td>
<td>22.9</td>
<td>2.9</td>
<td>9 (18-27)</td>
<td>0.13</td>
</tr>
<tr>
<td>Scenario 2</td>
<td></td>
<td>16.1</td>
<td>2.7</td>
<td>10 (10-20)</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>Checklist Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1</td>
<td>0-27</td>
<td>20.4</td>
<td>2.7</td>
<td>8 (16-24)</td>
<td>0.09</td>
</tr>
<tr>
<td>Scenario 2</td>
<td></td>
<td>14.7</td>
<td>2.1</td>
<td>8 (10-18)</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>GAS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1</td>
<td>0-3</td>
<td>2.5</td>
<td>0.5</td>
<td>1 (2-3)</td>
<td>0.00</td>
</tr>
<tr>
<td>Scenario 2</td>
<td></td>
<td>1.4</td>
<td>0.7</td>
<td>3 (0-3)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Figures 11.2 and 11.3, next page, demonstrate the distributions of total scores for each
simulation scenario individually. Results for the checklist components and total scores
for each simulation scenario, were normally distributed (Shapiro-Wilk significance >
0.05). Results for the GAS for each simulation scenario were not normally distributed
(Shapiro-Wilk significance < 0.05).
Figure 11.2: Distribution of total scores for Simulation Scenario 1

Figure 11.3: Distribution of total scores for Simulation Scenario 2
Results for each component of the simulation assessment tools were also analysed to determine the correlation between the checklist and GAS components, as shown in Table 11.4, Figure 11.4, and Figure 11.5. There was a significant (p < 0.05) correlation between the checklist and GAS components for Simulation Scenario 2, but not for Scenario 1.

Table 11.4: Correlation between the checklist and GAS components of the simulation assessment tools

<table>
<thead>
<tr>
<th>Simulation Scenario</th>
<th>N</th>
<th>Correlation</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>19</td>
<td>0.41</td>
<td>0.082</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>18</td>
<td>0.71</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Correlation: Checklist Score and GAS (Simulation Scenario 1)

Figure 11.4: Correlation between checklist score and GAS for Scenario 1
D. Correlation Between MCQ Test and Simulation Assessments

The results of the correlations between the MCQ test and simulation scenarios are presented in both numerical and graphical form, in Table 11.5, and Figures 11.6 and 11.7. There were statistically significant correlations (p < 0.05) between the MCQ scores and the total and component simulation scores, for both simulation scenarios. The strength of correlation varied slightly, but was approximately 0.5 for each assessment measure on each of the two scenarios.

Table 11.5: Correlation between the MCQ test and simulation assessments, including each component of each simulation assessment

<table>
<thead>
<tr>
<th></th>
<th>MCQ Test &amp; Scenario 1</th>
<th>MCQ Test &amp; Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation</td>
<td>p-value</td>
</tr>
<tr>
<td>Total</td>
<td>r = 0.52</td>
<td>0.023</td>
</tr>
<tr>
<td>Checklist</td>
<td>r = 0.47</td>
<td>0.044</td>
</tr>
<tr>
<td>GAS</td>
<td>ρ = 0.52</td>
<td>0.022</td>
</tr>
</tbody>
</table>
Figure 11.6: Correlation between the MCQ test and simulation assessments (Scenario 1)

Figure 11.7: Correlation between the MCQ test and simulation assessments (Scenario 2)
CHAPTER 12: CONTENT VALIDITY

A. Multiple Choice Question Test

Results from the qualitative analysis were tabulated for ease of comparison. Table 12.1 demonstrates the relationship of the MCQ items to the learning objective categories, for the purpose of analysing for adequate sampling. Each item is categorised by topic (AMI or APO) and learning outcome category (e.g. history, examination, etc.). The total number of items in each learning outcome category, for each topic, and for both topics combined, is tallied. This table demonstrates that each specific learning outcome category was covered by at least one assessment item (though not necessarily for each topic), indicating good content validity across MCQ assessment.

Table 12.2 shows the keywords identified for each item (2nd column) and then separates out those that were determined have relevance to either AMI or APO (3rd column). The purpose of this table was to analyse for irrelevant content. This table shows that each MCQ item has some relevance to either AMI or APO, which is indicative of good content validity.

It is important to note that there are some apparent discrepancies between the keywords listed for each of the MCQ items and the classification of each of the items into the learning outcome categories. This is because, in classifying the items by learning outcome, the focus was on identifying the competency that the item was directly assessing. This is in contrast to the process of allocating keywords, where the focus was on characterising any themes contained in the item/vignette.
Table 12.1: Learning outcomes for AMI and APO, and corresponding MCQ items

<table>
<thead>
<tr>
<th>Question</th>
<th>Topic</th>
<th>A. Patient assessment</th>
<th>B. Patient management</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>AMI</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AMI</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>AMI</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>AMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>AMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>AMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>AMI</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>18</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>AMI</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>APO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>APO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>APO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>APO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>APO</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>APO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>APO</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>30</td>
<td>APO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of AMI questions: 2, 0, 6, 3, 2, 5, 0
Number of APO questions: 0, 1, 5, 1, 1, 6, 2
Total number of questions: 2, 1, 11, 4, 3, 11, 2
Table 12.2: Keywords identified for each MCQ item

<table>
<thead>
<tr>
<th>Item</th>
<th>Keywords</th>
<th>Relevant Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TTE, investigation, IHD, CCF</td>
<td>IHD, CCF</td>
</tr>
<tr>
<td>2</td>
<td>Chest pain, investigation, management</td>
<td>Chest pain</td>
</tr>
<tr>
<td>3</td>
<td>Chest pain, investigation, ECG, AMI</td>
<td>Chest pain, AMI</td>
</tr>
<tr>
<td>4</td>
<td>Chest pain, AMI, management</td>
<td>Chest pain, AMI</td>
</tr>
<tr>
<td>5</td>
<td>Chest pain, investigation, ECG, AMI, management, prioritisation</td>
<td>Chest pain, AMI</td>
</tr>
<tr>
<td>6</td>
<td>Chest pain, AMI, management</td>
<td>Chest pain, AMI</td>
</tr>
<tr>
<td>7</td>
<td>Dyspnoea, ECG, ACS, diagnosis, management</td>
<td>Dyspnoea, ACS</td>
</tr>
<tr>
<td>8</td>
<td>ACS, management</td>
<td>ACS</td>
</tr>
<tr>
<td>9</td>
<td>ACS, investigation</td>
<td>ACS</td>
</tr>
<tr>
<td>10</td>
<td>ACS, management, prioritisation</td>
<td>ACS</td>
</tr>
<tr>
<td>11</td>
<td>IHD, history</td>
<td>IHD</td>
</tr>
<tr>
<td>12</td>
<td>Dyspnoea, CCF, HF, investigation</td>
<td>Dyspnoea, CCF, HF</td>
</tr>
<tr>
<td>13</td>
<td>HF, investigation</td>
<td>HF</td>
</tr>
<tr>
<td>14</td>
<td>HF, investigation</td>
<td>HF</td>
</tr>
<tr>
<td>15</td>
<td>Chest pain, AMI, history, management</td>
<td>Chest pain, AMI</td>
</tr>
<tr>
<td>16</td>
<td>AMI, management</td>
<td>AMI</td>
</tr>
<tr>
<td>17</td>
<td>AMI, ECG, investigation</td>
<td>AMI</td>
</tr>
<tr>
<td>18</td>
<td>Chest pain, AMI, ECG, investigation, management, prioritisation</td>
<td>Chest pain, AMI</td>
</tr>
<tr>
<td>19</td>
<td>IHD, ECG, investigation</td>
<td>IHD</td>
</tr>
<tr>
<td>20</td>
<td>Dyspnoea, APO, management</td>
<td>Dyspnoea, APO</td>
</tr>
<tr>
<td>21</td>
<td>CCF, HF, management</td>
<td>CCF, HF</td>
</tr>
<tr>
<td>22</td>
<td>AMI, APO, diagnosis</td>
<td>AMI, APO</td>
</tr>
<tr>
<td>23</td>
<td>APO, examination, dyspnoea</td>
<td>APO</td>
</tr>
<tr>
<td>24</td>
<td>APO, examination, patient progress</td>
<td>APO</td>
</tr>
<tr>
<td>25</td>
<td>APO, investigation</td>
<td>APO</td>
</tr>
<tr>
<td>26</td>
<td>IHD, CCF, APO, dyspnoea, management</td>
<td>IHD, CCF, APO, dyspnoea</td>
</tr>
<tr>
<td>27</td>
<td>APO, ECG, investigation, AF, management</td>
<td>APO</td>
</tr>
<tr>
<td>28</td>
<td>APO, examination, management</td>
<td>APO</td>
</tr>
<tr>
<td>29</td>
<td>APO, investigation, management, prioritization</td>
<td>APO</td>
</tr>
<tr>
<td>30</td>
<td>APO, management</td>
<td>APO</td>
</tr>
</tbody>
</table>
B. Simulation Assessments

i. Simulation Scenario 1

Table 12.3 demonstrates the relationship of the Simulation Scenario 1 assessment items to the learning outcome categories (e.g. history, examination, etc.). This table demonstrates that each specific learning outcome category was covered by at least one assessment item in Simulation Scenario 1, indicating good content validity across this assessment tool.

**Table 12.3: Learning outcome categories for AMI and corresponding simulation assessment items (Scenario 1)**

<table>
<thead>
<tr>
<th>Item</th>
<th>1. Patient Assessment</th>
<th>B. Patient Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A1b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A1c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A1d</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A1e</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A3a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A4a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A5a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A5b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A5c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A5d</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A5e</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A6a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A6b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A6c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A6d</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1d</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>B2a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B2b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B3a</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

| Number of items | 10 | 5 | 3 | 1 | 4 | 3 | 1 |
Table 12.4 demonstrates the relationship of the Scenario 1 items to specific learning outcomes. The specific learning outcomes are listed in the column on the left, and the corresponding simulation assessment items are listed in the column on the right. Again, each learning outcome is covered by at least one assessment item, indicating good content validity.

**Table 12.4: Mapping of simulation assessment items (Scenario 1) to specific learning outcomes**

<table>
<thead>
<tr>
<th>Specific Learning Outcome</th>
<th>Simulation Assessment Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Takes a focused history to differentiate between causes of chest pain</td>
<td>A1a, A1b, A1c, A1d, A1e</td>
</tr>
<tr>
<td></td>
<td>A2a, A2b, A2c</td>
</tr>
<tr>
<td></td>
<td>A3a, A4a</td>
</tr>
<tr>
<td>Performs a focused cardiorespiratory examination</td>
<td>A5a, A5b, A5c, A5d, A5e</td>
</tr>
<tr>
<td>Requests appropriate investigations</td>
<td>A6a, A6c, A6d</td>
</tr>
<tr>
<td>Interprets an ECG</td>
<td>A6b*</td>
</tr>
<tr>
<td>Recognises patients who are acutely ill and require immediate treatment</td>
<td>A6b*, B3a*</td>
</tr>
<tr>
<td>Recognises when to call for help</td>
<td>B3a*</td>
</tr>
<tr>
<td>Is able to form a list of differential diagnoses and determine the most likely diagnosis</td>
<td>A6b*</td>
</tr>
<tr>
<td>Identifies and initiates appropriate initial management for acute coronary syndrome</td>
<td>B1a, B1b, B1c, B1d</td>
</tr>
<tr>
<td>Identifies the need for senior clinician input to determine and implement further management</td>
<td>B3a*</td>
</tr>
<tr>
<td>Re-evaluates the patient’s progress and response to treatment</td>
<td>B2a, B2b</td>
</tr>
</tbody>
</table>
ii. Simulation Scenario 2

Table 12.5, below, demonstrates the relationship of the simulation Scenario 2 assessment items to the learning outcome categories. The total number of items in each learning outcome category is tallied.

Table 12.5: Learning outcome categories for APO and corresponding simulation assessment items (Scenario 2)

<table>
<thead>
<tr>
<th>Item</th>
<th>1. Patient Assessment</th>
<th>B. Patient Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A1b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2d</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2e</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2f</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A2g</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A3a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A3b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A3c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A3d</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A3e</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A4a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A4b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A4c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A4d</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>A4e</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1c</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1d</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B1e</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B2a</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B2b</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>B3a</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Number of items | 9 | 5 | 4 | 1 | 5 | 2 | 1
Table 12.6 demonstrates the relationship of the Scenario 2 items to specific learning outcomes. The specific learning outcomes are listed in the column on the left, and the corresponding simulation assessment items are listed in the column on the right. As for Simulation Scenario 1, each learning outcome category, and each specific learning outcome, is assessed by at least one item, suggesting good content validity.

**Table 12.6: Mapping of simulation assessment items (Scenario 2) to specific learning outcomes**

* Items A4b, B1e and B3a assess more than one learning outcome

<table>
<thead>
<tr>
<th>Specific Learning Outcome</th>
<th>Simulation Assessment Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Takes a focused history to differentiate between causes of dyspnoea</td>
<td>A1a, A1b</td>
</tr>
<tr>
<td></td>
<td>A2a, A2b, A2c, A2d, A2e, A2f, A2g</td>
</tr>
<tr>
<td>Performs a focused cardiorespiratory examination</td>
<td>A3a, A3b, A3c, A3d, A3e</td>
</tr>
<tr>
<td>Requests appropriate investigations</td>
<td>A4a, A4c, A4d, A4e</td>
</tr>
<tr>
<td>Interprets an ECG</td>
<td>A4b*</td>
</tr>
<tr>
<td>Recognises patients who are acutely ill and require immediate treatment</td>
<td>A4b*</td>
</tr>
<tr>
<td></td>
<td>B3a*</td>
</tr>
<tr>
<td>Recognises when to call for help</td>
<td>B3a*</td>
</tr>
<tr>
<td>Is able to form a list of differential diagnoses and determine the most likely diagnosis</td>
<td>A4b</td>
</tr>
<tr>
<td></td>
<td>B1e*</td>
</tr>
<tr>
<td>Identifies and initiates appropriate initial management for acute pulmonary oedema/Fluid overload</td>
<td>B1a, B1b, B1c, B1d, B1e*</td>
</tr>
<tr>
<td>Identifies the need for senior clinician input to determine and implement further management</td>
<td>B3a*</td>
</tr>
<tr>
<td>Re-evaluates the patient’s progress and response to treatment</td>
<td>B2a, B2b</td>
</tr>
</tbody>
</table>
CHAPTER 13: CONSTRUCT VALIDITY

A. Multiple Choice Question Test

i. Internal Consistency

Table 13.1, below, shows internal consistency statistics for the MCQ test, when considered as a single assessment tool, and when divided into two separate assessment tools (divided into AMI and APO topics). Coefficient alpha scores all fail to reach an acceptable level, but are particularly low for the APO items, where there is essentially no correlation between items.

Table 13.1: Internal consistency statistics for the MCQ test
* Note: Item 20 was excluded due to zero variance for this item

<table>
<thead>
<tr>
<th></th>
<th>No. of items</th>
<th>Mean</th>
<th>SD</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overall</td>
</tr>
<tr>
<td><strong>All items</strong></td>
<td>29</td>
<td>21.9</td>
<td>3.4</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>AMI items</strong></td>
<td>15</td>
<td>10.9</td>
<td>2.8</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>APO items</strong></td>
<td>14</td>
<td>11.1</td>
<td>1.4</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table 13.2 shows Cronbach’s alpha for each MCQ item, when the entire MCQ test is considered as a single assessment tool. This shows which items best correlate with the other items on the assessment (higher corrected item-total correlation), and which items might detract from the overall internal consistency (higher Cronbach’s alpha if item deleted).

**Table 13.2: Cronbach’s alpha for each MCQ item (all items considered as a single assessment tool)**

<table>
<thead>
<tr>
<th>Item</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.28</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>-0.13</td>
<td>0.65</td>
</tr>
<tr>
<td>3</td>
<td>0.26</td>
<td>0.61</td>
</tr>
<tr>
<td>4</td>
<td>0.28</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>0.45</td>
<td>0.59</td>
</tr>
<tr>
<td>6</td>
<td>0.45</td>
<td>0.60</td>
</tr>
<tr>
<td>7</td>
<td>0.42</td>
<td>0.61</td>
</tr>
<tr>
<td>8</td>
<td>0.30</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>0.39</td>
<td>0.60</td>
</tr>
<tr>
<td>10</td>
<td>0.29</td>
<td>0.61</td>
</tr>
<tr>
<td>11</td>
<td>0.42</td>
<td>0.59</td>
</tr>
<tr>
<td>12</td>
<td>-0.13</td>
<td>0.65</td>
</tr>
<tr>
<td>13</td>
<td>0.39</td>
<td>0.60</td>
</tr>
<tr>
<td>14</td>
<td>0.33</td>
<td>0.61</td>
</tr>
<tr>
<td>15</td>
<td>-0.03</td>
<td>0.63</td>
</tr>
<tr>
<td>16</td>
<td>0.23</td>
<td>0.61</td>
</tr>
<tr>
<td>17</td>
<td>0.24</td>
<td>0.61</td>
</tr>
<tr>
<td>18</td>
<td>0.29</td>
<td>0.61</td>
</tr>
<tr>
<td>19</td>
<td>0.43</td>
<td>0.59</td>
</tr>
<tr>
<td>20</td>
<td>-0.05</td>
<td>0.63</td>
</tr>
<tr>
<td>21</td>
<td>-0.06</td>
<td>0.64</td>
</tr>
<tr>
<td>22</td>
<td>0.13</td>
<td>0.62</td>
</tr>
<tr>
<td>23</td>
<td>0.21</td>
<td>0.62</td>
</tr>
<tr>
<td>24</td>
<td>0.29</td>
<td>0.62</td>
</tr>
<tr>
<td>25</td>
<td>-0.10</td>
<td>0.65</td>
</tr>
<tr>
<td>26</td>
<td>-0.07</td>
<td>0.65</td>
</tr>
<tr>
<td>27</td>
<td>0.05</td>
<td>0.63</td>
</tr>
<tr>
<td>28</td>
<td>0.09</td>
<td>0.63</td>
</tr>
<tr>
<td>29</td>
<td>-0.06</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Table 13.3 shows Cronbach’s alpha for each item on the MCQ test when items are divided by topic (AMI or APO), i.e. the separate sections are considered as two separate assessment tools. The interpretation of this table is as for Table 13.2.

Table 13.3: Cronbach’s alpha for each item on the MCQ test (items divided by topic and considered as two separate assessment tools)

<table>
<thead>
<tr>
<th>Item</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AMI Items</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.70</td>
</tr>
<tr>
<td>3</td>
<td>0.16</td>
<td>0.68</td>
</tr>
<tr>
<td>4</td>
<td>0.26</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>0.37</td>
<td>0.66</td>
</tr>
<tr>
<td>6</td>
<td>0.45</td>
<td>0.66</td>
</tr>
<tr>
<td>7</td>
<td>0.37</td>
<td>0.67</td>
</tr>
<tr>
<td>8</td>
<td>0.36</td>
<td>0.66</td>
</tr>
<tr>
<td>9</td>
<td>0.36</td>
<td>0.66</td>
</tr>
<tr>
<td>10</td>
<td>0.33</td>
<td>0.66</td>
</tr>
<tr>
<td>11</td>
<td>0.37</td>
<td>0.66</td>
</tr>
<tr>
<td>15</td>
<td>-0.01</td>
<td>0.70</td>
</tr>
<tr>
<td>16</td>
<td>0.32</td>
<td>0.66</td>
</tr>
<tr>
<td>17</td>
<td>0.33</td>
<td>0.66</td>
</tr>
<tr>
<td>18</td>
<td>0.23</td>
<td>0.68</td>
</tr>
<tr>
<td>19</td>
<td>0.47</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>APO Items</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.29</td>
<td>-0.20</td>
</tr>
<tr>
<td>12</td>
<td>-0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>13</td>
<td>0.25</td>
<td>-0.09</td>
</tr>
<tr>
<td>14</td>
<td>0.26</td>
<td>-0.07</td>
</tr>
<tr>
<td>21</td>
<td>-0.10</td>
<td>0.07</td>
</tr>
<tr>
<td>22</td>
<td>-0.32</td>
<td>0.19</td>
</tr>
<tr>
<td>23</td>
<td>0.18</td>
<td>-0.06</td>
</tr>
<tr>
<td>24</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>25</td>
<td>0.13</td>
<td>0.01</td>
</tr>
<tr>
<td>26</td>
<td>-0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>27</td>
<td>-0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>28</td>
<td>0.13</td>
<td>0.01</td>
</tr>
<tr>
<td>29</td>
<td>0.17</td>
<td>-0.10</td>
</tr>
<tr>
<td>30</td>
<td>0.01</td>
<td>0.04</td>
</tr>
</tbody>
</table>
ii. Contrasted-Groups Approach

Figure 13.1, below, demonstrates the MCQ test results of the most junior (MD2) and most senior (HMO2) participants, representing novice and expert groups, respectively. There is a statistically significant difference (p < 0.05) between MCQ scores for these two classifications.

Figure 13.1: Boxplot comparing MCQ scores for MD2 and HMO2 groups
MD2: n = 8, mean (SD) = 22.3 (3.7), median = 23.0
HMO2: n = 11, mean (SD) = 25.8 (2.1), median = 25.0
Mann-Whitney p-value = 0.041
iii. Comparison Across All Groups

Figure 13.2, below, demonstrates the MCQ test results of participants across all classifications, from most junior to most senior. Table 13.4 shows the Kruskal-Wallis comparison pairwise across all classifications, obtained using the Dunn-Bonferroni post-hoc method. Figure 13.2 shows that the MCQ score differs significantly ($p < 0.05$) with classification, but Table 13.4 shows that the only significant difference ($p < 0.05$) between any two groups is between the intern and HMO2 groups.

![Figure 13.2: Boxplot comparing MCQ scores across all classifications](image)

MD2: $n = 8$, mean (SD) = 22.3 (3.7), median = 23.0  
MD3: $n = 6$, mean (SD) = 24.1 (1.9), median = 23.5  
MD4: $n = 3$, mean (SD) = 21.0 (4.6), median = 22.0  
Intern: $n = 10$, mean (SD) = 20.2 (2.3), median = 20.0  
HMO2: $n = 11$, mean (SD) = 25.8 (2.1), median = 25.0  
Kruskal-Wallis p-value = 0.002
Table 13.4: Kruskal-Wallis comparison pairwise across all classifications for MCQ scores

<table>
<thead>
<tr>
<th>Classifications Compared</th>
<th>Test Statistic</th>
<th>Adj. Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMO2 - Intern</td>
<td>18.8</td>
<td>0.001</td>
</tr>
<tr>
<td>HMO2 - MD4</td>
<td>15.0</td>
<td>0.379</td>
</tr>
<tr>
<td>HMO2 - MD3</td>
<td>6.2</td>
<td>1.000</td>
</tr>
<tr>
<td>HMO2 - MD2</td>
<td>11.2</td>
<td>0.292</td>
</tr>
<tr>
<td>Intern - MD4</td>
<td>3.8</td>
<td>1.000</td>
</tr>
<tr>
<td>Intern - MD3</td>
<td>12.6</td>
<td>0.280</td>
</tr>
<tr>
<td>Intern - MD2</td>
<td>7.6</td>
<td>1.000</td>
</tr>
<tr>
<td>MD4 - MD3</td>
<td>8.8</td>
<td>1.000</td>
</tr>
<tr>
<td>MD4 - MD2</td>
<td>3.8</td>
<td>1.000</td>
</tr>
<tr>
<td>MD3 - MD2</td>
<td>5.0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

B. Simulation Assessments

i. Internal Consistency

Table 13.5, below, shows internal consistency statistics for the simulation assessments.

The coefficient alpha scores both fail to reach an acceptable level, but is particularly low for Scenario 2, where there is essentially no correlation between items.

Table 13.5: Internal consistency statistics for the simulation assessments

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Cronbach’s Alpha Overall</th>
<th>Standardised Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>0.46</td>
<td>0.52</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>-0.06</td>
<td>-0.03</td>
</tr>
</tbody>
</table>
Tables 13.6 and 13.7 show Cronbach’s alpha for each item on each simulation scenario.

In calculating internal consistency statistics, those items with no variance are removed from the scale: three items in Scenario 1 (A1e, A5a, A6a), and two items in Scenario 2 (A1b, A3d) are not included in calculations of internal consistency for this reason.

These tables show which items best correlate with the other items on the assessment (higher corrected item-total correlation), and which items might detract from the overall internal consistency (higher Cronbach’s alpha if item deleted).

**Table 13.6: Cronbach’s alpha for each item on Simulation 1**

Note: Scenario 1 item a has a maximum attainable mark of 2

<table>
<thead>
<tr>
<th>Item</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1a</td>
<td>0.06</td>
<td>0.46</td>
</tr>
<tr>
<td>A1b</td>
<td>0.32</td>
<td>0.44</td>
</tr>
<tr>
<td>A1c</td>
<td>-0.08</td>
<td>0.50</td>
</tr>
<tr>
<td>A1d</td>
<td>-0.07</td>
<td>0.48</td>
</tr>
<tr>
<td>A2a</td>
<td>0.12</td>
<td>0.45</td>
</tr>
<tr>
<td>A2b</td>
<td>0.24</td>
<td>0.43</td>
</tr>
<tr>
<td>A2c</td>
<td>-0.04</td>
<td>0.49</td>
</tr>
<tr>
<td>A3a</td>
<td>-0.19</td>
<td>0.56</td>
</tr>
<tr>
<td>A4</td>
<td>0.49</td>
<td>0.36</td>
</tr>
<tr>
<td>A5b</td>
<td>0.28</td>
<td>0.42</td>
</tr>
<tr>
<td>A5c</td>
<td>0.06</td>
<td>0.46</td>
</tr>
<tr>
<td>A5d</td>
<td>-0.08</td>
<td>0.49</td>
</tr>
<tr>
<td>A5e</td>
<td>0.17</td>
<td>0.44</td>
</tr>
<tr>
<td>A6b</td>
<td>-0.14</td>
<td>0.48</td>
</tr>
<tr>
<td>A6c</td>
<td>-0.19</td>
<td>0.50</td>
</tr>
<tr>
<td>A6d</td>
<td>0.03</td>
<td>0.47</td>
</tr>
<tr>
<td>B1a</td>
<td>0.33</td>
<td>0.42</td>
</tr>
<tr>
<td>B1b</td>
<td>-0.03</td>
<td>0.48</td>
</tr>
<tr>
<td>B1c</td>
<td>0.32</td>
<td>0.44</td>
</tr>
<tr>
<td>B1d</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>B2a</td>
<td>0.59</td>
<td>0.35</td>
</tr>
<tr>
<td>B2b</td>
<td>0.39</td>
<td>0.41</td>
</tr>
<tr>
<td>B3a</td>
<td>0.45</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Table 13.7: Cronbach’s alpha for each item on Simulation Scenario 2

<table>
<thead>
<tr>
<th>Item</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach's Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1a</td>
<td>-0.52</td>
<td>0.18</td>
</tr>
<tr>
<td>A2a</td>
<td>-0.44</td>
<td>0.015</td>
</tr>
<tr>
<td>A2b</td>
<td>0.47</td>
<td>-0.18</td>
</tr>
<tr>
<td>A2c</td>
<td>-0.05</td>
<td>-0.04</td>
</tr>
<tr>
<td>A2d</td>
<td>-0.38</td>
<td>0.12</td>
</tr>
<tr>
<td>A2e</td>
<td>-0.30</td>
<td>0.07</td>
</tr>
<tr>
<td>A2f</td>
<td>-0.49</td>
<td>0.10</td>
</tr>
<tr>
<td>A2g</td>
<td>-0.30</td>
<td>0.01</td>
</tr>
<tr>
<td>A3a</td>
<td>-0.03</td>
<td>-0.05</td>
</tr>
<tr>
<td>A3b</td>
<td>0.17</td>
<td>-0.16</td>
</tr>
<tr>
<td>A3c</td>
<td>0.05</td>
<td>-0.08</td>
</tr>
<tr>
<td>A3e</td>
<td>0.27</td>
<td>-0.18</td>
</tr>
<tr>
<td>A4a</td>
<td>0.54</td>
<td>-0.36</td>
</tr>
<tr>
<td>A4b</td>
<td>0.18</td>
<td>-0.14</td>
</tr>
<tr>
<td>A4c</td>
<td>0.70</td>
<td>-0.52</td>
</tr>
<tr>
<td>A4d</td>
<td>0.49</td>
<td>-0.33</td>
</tr>
<tr>
<td>A4e</td>
<td>0.29</td>
<td>-0.13</td>
</tr>
<tr>
<td>B1a</td>
<td>-0.24</td>
<td>0.06</td>
</tr>
<tr>
<td>B1b</td>
<td>-0.39</td>
<td>0.11</td>
</tr>
<tr>
<td>B1c</td>
<td>-0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>B1d</td>
<td>0.15</td>
<td>-0.15</td>
</tr>
<tr>
<td>B1e</td>
<td>0.20</td>
<td>-0.18</td>
</tr>
<tr>
<td>B2a</td>
<td>0.02</td>
<td>-0.07</td>
</tr>
<tr>
<td>B2b</td>
<td>0.11</td>
<td>-0.11</td>
</tr>
<tr>
<td>B3a</td>
<td>0.09</td>
<td>-0.11</td>
</tr>
</tbody>
</table>
ii. Contrasted-Groups Approach

Figure 13.3 below, demonstrates the simulation assessment results (total score) of the most junior and most senior participants for Scenario 1. There is no significant difference in the scores for these two groups (p > 0.05).

Figure 13.3: Boxplot comparing simulation assessment scores for Scenario 1 for MD3 and HMO2 groups
MD3: n = 4, mean (SD) = 23.3 (2.2), median = 24.0
HMO2: n = 5, mean (SD) = 24.4 (2.7), median = 25.0
Mann-Whitney p-value = 0.315
Figure 13.4 below, demonstrates the simulation assessment results (total score) of the most junior and most senior participants for Simulation Scenario 1. There is a statistically significant difference in the scores for these two groups (p < 0.05).

Figure 13.4: Boxplot comparing simulation assessment scores for Scenario 2 for MD2 and HMO2 groups
MD2: n = 7, mean (SD) = 13.4 (1.7), median = 14.0
HMO2: n = 6, mean (SD) = 18.2 (1.9), median = 18.5
Mann-Whitney p-value = 0.004
Table 13.8, below, demonstrates the simulation assessment results (simulation components- checklist score and GAS) of the most junior and most senior participants. There is no statistically significant difference between the novice and expert groups (p > 0.05) for either component of Scenario 1, but the difference is significant (p < 0.05) for both the checklist and GAS components for Scenario 2.

Table 13.8: Comparison of simulation component scores for the ‘novice’ and ‘expert’ groups
Note: Only one MD2 participant completed Simulation Scenario 1, so the MD3 group was used as the ‘novice’ group for Scenario 1

<table>
<thead>
<tr>
<th>Simulation Scenario</th>
<th>Class</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Mann-Whitney p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Checklist Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1</td>
<td>MD3*</td>
<td>4</td>
<td>20.5</td>
<td>2.4</td>
<td>0.381</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>21.6</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>MD2</td>
<td>7</td>
<td>12.6</td>
<td>1.4</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>16.2</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td><strong>GAS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1</td>
<td>MD3*</td>
<td>4</td>
<td>2.8</td>
<td>0.5</td>
<td>0.866</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>2.8</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>MD2</td>
<td>7</td>
<td>0.9</td>
<td>0.4</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>2.0</td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>
iii. Comparison Across All Groups

The results of the independent samples Kruskal-Wallis test, which was performed to analyse for differences across all classifications for the total score, checklist score and GAS components of each simulation scenario, are shown in Figures 13.5 and 13.6, and Table 13.9, below. The total score, checklist score, and GAS component demonstrated significant differences (p < 0.05) across classifications for Scenario 2, but not for Scenario 1.

![Boxplot comparing simulation assessment scores across all classifications for Simulation Scenario 1 (total score)](image)

**Figure 13.5: Boxplot comparing simulation assessment scores across all classifications for Simulation Scenario 1 (total score)**

MD2: n = 1, mean = 20.0*, median = 20.0*
MD3: n = 4, mean (SD) = 23.3 (2.2), median = 24.0
MD4: n = 3, mean (SD) = 20.7 (2.5), median = 21.0
Intern: n = 6, mean (SD) = 23.0 (2.6), median = 23.5
HMO2: n = 5, mean (SD) = 24.4 (2.7), median = 25.0

Kruskal-Wallis p-value = 0.399

*Note: For Scenario 1, the MD2 results represent a single participant’s score, not a mean score for MD2 classification, as only one MD2 participant completed Simulation Scenario 1.
Figure 13.6: Boxplot comparing simulation assessment scores across all classifications for Simulation Scenario 2 (total score)
MD2: n = 7, mean (SD) = 13.4 (1.7), median = 14.0
MD3: n = 2, mean (SD) = 16.5 (0.7), median = 16.5
Intern: n = 3, mean (SD) = 17.7 (0.6), median = 18.0
HMO2: n = 6, mean (SD) = 18.2 (1.9), median = 18.5
Kruskal-Wallis p-value = 0.006
Table 13.9: Comparison of simulation component scores across all classifications

*Note: For Scenario 1, the MD2 results represent a single participant’s score, not a mean score for MD2 classification, as only one MD2 participant completed Simulation Scenario 1.

<table>
<thead>
<tr>
<th>Simulation Scenario</th>
<th>Class</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>Kruskal-Wallis p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Checklist Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1</td>
<td>MD2</td>
<td>1</td>
<td>18.0*</td>
<td>.</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>4</td>
<td>20.5</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD4</td>
<td>3</td>
<td>18.3</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>6</td>
<td>20.7</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>21.6</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>19</td>
<td>20.4</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>MD2</td>
<td>7</td>
<td>12.6</td>
<td>1.4</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>2</td>
<td>15.0</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD4</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>3</td>
<td>16.3</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>16.2</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18</td>
<td>14.7</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td><strong>GAS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 1</td>
<td>MD2</td>
<td>1</td>
<td>2.0*</td>
<td>.</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>4</td>
<td>2.8</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD4</td>
<td>3</td>
<td>2.3</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>6</td>
<td>2.3</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>2.8</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>19</td>
<td>2.5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>MD2</td>
<td>7</td>
<td>0.9</td>
<td>0.4</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>2</td>
<td>1.5</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD4</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>3</td>
<td>1.3</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>2.0</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18</td>
<td>1.4</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>

Pairwise comparison across all classifications for each of the total score, checklist component and GAS component for Scenario 2 demonstrated significant differences (adjusted significance value < 0.05) between the MD2 and HMO2 groups only. There were no significant differences between the performance of other groups on Scenario 2, in terms of total score, checklist score or GAS component.
CHAPTER 14: FORMATIVE VALIDITY

A. Multiple Choice Question Test

i. Quantitative Component

Table 14.1 shows the mean item value (difficulty) and standard deviation for each item of the MCQ test. Medium item difficulty for an MCQ test with three options per item is approximately 0.67. (98)

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean Item Value</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.63</td>
<td>0.49</td>
</tr>
<tr>
<td>2</td>
<td>0.87</td>
<td>0.34</td>
</tr>
<tr>
<td>3</td>
<td>0.79</td>
<td>0.41</td>
</tr>
<tr>
<td>4</td>
<td>0.63</td>
<td>0.49</td>
</tr>
<tr>
<td>5</td>
<td>0.79</td>
<td>0.41</td>
</tr>
<tr>
<td>6</td>
<td>0.92</td>
<td>0.27</td>
</tr>
<tr>
<td>7</td>
<td>0.92</td>
<td>0.27</td>
</tr>
<tr>
<td>8</td>
<td>0.45</td>
<td>0.50</td>
</tr>
<tr>
<td>9</td>
<td>0.87</td>
<td>0.34</td>
</tr>
<tr>
<td>10</td>
<td>0.63</td>
<td>0.49</td>
</tr>
<tr>
<td>11</td>
<td>0.66</td>
<td>0.48</td>
</tr>
<tr>
<td>12</td>
<td>0.87</td>
<td>0.34</td>
</tr>
<tr>
<td>13</td>
<td>0.87</td>
<td>0.34</td>
</tr>
<tr>
<td>14</td>
<td>0.92</td>
<td>0.27</td>
</tr>
<tr>
<td>15</td>
<td>0.92</td>
<td>0.27</td>
</tr>
<tr>
<td>16</td>
<td>0.66</td>
<td>0.48</td>
</tr>
<tr>
<td>17</td>
<td>0.58</td>
<td>0.50</td>
</tr>
<tr>
<td>18</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td>19</td>
<td>0.58</td>
<td>0.50</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>21</td>
<td>0.97</td>
<td>0.16</td>
</tr>
<tr>
<td>22</td>
<td>0.87</td>
<td>0.34</td>
</tr>
<tr>
<td>23</td>
<td>0.87</td>
<td>0.34</td>
</tr>
<tr>
<td>24</td>
<td>0.92</td>
<td>0.27</td>
</tr>
<tr>
<td>25</td>
<td>0.97</td>
<td>0.16</td>
</tr>
<tr>
<td>26</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>27</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>28</td>
<td>0.97</td>
<td>0.16</td>
</tr>
<tr>
<td>29</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>30</td>
<td>0.63</td>
<td>0.49</td>
</tr>
<tr>
<td>Total</td>
<td>22.95</td>
<td>3.45</td>
</tr>
</tbody>
</table>
Table 14.2 and Figure 14.1 show the spread of MCQ item difficulty values. There were relatively more items that were easy (numerically higher item difficulty - 0.91-1.0) and fewer items that were very challenging (item difficulty < 0.5). An alternative interpretation is that the assessment better discriminates between lower performing participants.

**Table 14.2: Spread of item difficulty values for the MCQ test**

<table>
<thead>
<tr>
<th>Item Difficulty</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.41-0.50</td>
<td>1</td>
</tr>
<tr>
<td>0.51-0.60</td>
<td>5</td>
</tr>
<tr>
<td>0.61-0.70</td>
<td>7</td>
</tr>
<tr>
<td>0.71-0.80</td>
<td>2</td>
</tr>
<tr>
<td>0.81-0.90</td>
<td>6</td>
</tr>
<tr>
<td>0.91-1.0</td>
<td>9</td>
</tr>
</tbody>
</table>

**Figure 14.1: Spread of item difficulty values for the MCQ test**

Table 14.3 (next page) shows the topics and categories of learning outcomes covered, sorted by difficulty. This table demonstrates that, for all learning outcome categories that were assessed by more than one item, the items were of different levels of difficulty, thereby allowing differentiation of candidates’ performance in each of these learning outcome categories.
14.3: Topics and categories of learning outcomes covered by MCQ items, sorted by difficulty

Note: * Items 3, 5, 17 and 29 fall under two learning outcome categories

<table>
<thead>
<tr>
<th>Learning Outcome Category</th>
<th>ITEM DIFFICULTY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.41-0.50</td>
</tr>
<tr>
<td>AMI ITEMS (Item Numbers)</td>
<td></td>
</tr>
<tr>
<td>A. Patient Assessment</td>
<td></td>
</tr>
<tr>
<td>1. History</td>
<td>11</td>
</tr>
<tr>
<td>2. Examination</td>
<td></td>
</tr>
<tr>
<td>3. Investigation</td>
<td>17*, 19</td>
</tr>
<tr>
<td>4. Diagnosis</td>
<td>17*</td>
</tr>
<tr>
<td>5. Prioritisation</td>
<td>10</td>
</tr>
<tr>
<td>B. Patient management</td>
<td></td>
</tr>
<tr>
<td>1. Management</td>
<td>8</td>
</tr>
<tr>
<td>2. Patient progress</td>
<td></td>
</tr>
<tr>
<td>APO ITEMS (Item Numbers)</td>
<td></td>
</tr>
<tr>
<td>A. Patient Assessment</td>
<td></td>
</tr>
<tr>
<td>1. History</td>
<td></td>
</tr>
<tr>
<td>2. Examination</td>
<td></td>
</tr>
<tr>
<td>3. Investigation</td>
<td>1</td>
</tr>
<tr>
<td>4. Diagnosis</td>
<td></td>
</tr>
<tr>
<td>5. Prioritisation</td>
<td></td>
</tr>
<tr>
<td>B. Patient management</td>
<td></td>
</tr>
<tr>
<td>1. Management</td>
<td>29*, 26, 27</td>
</tr>
<tr>
<td>2. Patient progress</td>
<td>30</td>
</tr>
</tbody>
</table>

There were three items on which MD2 participants performed better than HMO2s, seven items on which there was no difference in mean performance between the two groups, and 21 items on which HMO2 participants performed better than MD2s. A Mann-Whitney test comparing performance of MD2 and HMO2 participants on each MCQ item demonstrated that two MCQ items (items 8 and 9) had statistically significant differences in results for MD2 and HMO2 groups (p < 0.05). Results for these items are shown in Table 14.4 (next page).
### Table 14.4: Results for MD2 and HMO2 participants for MCQ items 8 and 9

<table>
<thead>
<tr>
<th>Item</th>
<th>Class</th>
<th>n</th>
<th>Mean Item Value</th>
<th>Mann-Whitney p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>MD2</td>
<td>8</td>
<td>0.13</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>11</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MD2</td>
<td>8</td>
<td>0.63</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>11</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

A Kruskal-Wallis test comparing performance across all classifications (MD2 to HMO2) demonstrated that one MCQ item (item 25- an item on the topic of initial investigations for APO) had statistically significant differences in results across the classifications. Results for this item are shown in Tables 14.5 and 14.6.

### Table 14.5: Results for participants of all classifications for MCQ item 25

<table>
<thead>
<tr>
<th>Item</th>
<th>Class</th>
<th>n</th>
<th>Mean Item Value</th>
<th>Kruskal-Wallis p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>MD2</td>
<td>8</td>
<td>1.00</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>6</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD4</td>
<td>3</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>10</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>11</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>38</td>
<td>0.97</td>
<td></td>
</tr>
</tbody>
</table>

### Table 14.6: Kruskal-Wallis comparison pairwise across all classifications for MCQ Item 25

<table>
<thead>
<tr>
<th>Classifications Compared</th>
<th>Test Statistic</th>
<th>Adj. Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMO2 - Intern</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>HMO2 - MD4</td>
<td>6.33</td>
<td>0.016</td>
</tr>
<tr>
<td>HMO2 - MD3</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>HMO2 - MD2</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>Intern - MD4</td>
<td>6.33</td>
<td>0.018</td>
</tr>
<tr>
<td>Intern - MD3</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>Intern - MD2</td>
<td>0.00</td>
<td>1.000</td>
</tr>
<tr>
<td>MD4 - MD3</td>
<td>6.33</td>
<td>0.037</td>
</tr>
<tr>
<td>MD4 - MD2</td>
<td>6.33</td>
<td>0.024</td>
</tr>
<tr>
<td>MD3- MD2</td>
<td>0.00</td>
<td>1.000</td>
</tr>
</tbody>
</table>
ii. Qualitative Component

Figure 14.2, below, shows the findings of the qualitative analysis of the MCQ feedback document. The four sections of the figure show the domains of feedback that are assessed in the evaluation of formative validity. These are: form, type, quantity and quality. Within each of these domains are descriptive findings of the qualitative analysis, which show the characteristics of the feedback provided to participants in regards to the MCQ test.

![Feedback Domains Diagram]

**Figure 14.2: Results of qualitative analysis of the MCQ Test Feedback document**

The form of the feedback document is a PDF document, which was sent to participants via email. The document contains each of the MCQ questions, in their original form, accompanied by answers and explanations for the answers. Individual participant marks were not given. In regards to the type of feedback, the document was designed to provide unidirectional group feedback; there was no individualisation of the feedback document for different participants and given that the document was sent via email,
there was no capacity for face-to-face contact to allow for bidirectional discussion (such as participants asking questions or seeking clarification). There was a specific focus on explaining the clinical context of the item stems, rather than focusing on the reason that the given answer was correct (though this was still explained).

The quantity of the feedback was two to three sentences per item, which amounted to a total nine-page document (compared to the original five-page MCQ test, four pages of answers/explanations were added). Evaluation of the quality of the feedback demonstrated that all information provided was factually correct, based on sound sources, and there were some references provided (which both shows academic rigor as well as providing an avenue for further learning).

While these descriptions do not offer any new information, in that they are essentially descriptions of the design of the MCQ test feedback document, they clearly show the properties of the feedback given. The implications of these findings are discussed in Chapter 20.

B. Simulation Assessments

i. Quantitative Component: Simulation Scenario 1

Table 14.7 shows the mean item value and standard deviation for each item of Simulation Scenario 1. In the context of a simulation assessment checklist, mean item values are a measure of the proportion of examinees satisfactorily executing a particular step. Unlike MCQ items, simulation items are not specifically designed to vary in difficulty, but rather represent the steps that should be taken to complete ‘ideal’ patient assessment and management. Those items with lower mean values represent steps in assessment/management that were not completed, and perhaps overlooked, by many
Table 14.7: Mean item values and standard deviation for each item of Simulation Scenario 1
Note: Item 3a has a maximum attainable mark of 2, therefore difficulty is half of the mean item value.

<table>
<thead>
<tr>
<th>Item/Component</th>
<th>Mean Item Value</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1a</td>
<td>0.89</td>
<td>0.32</td>
</tr>
<tr>
<td>A1b</td>
<td>0.95</td>
<td>0.23</td>
</tr>
<tr>
<td>A1c</td>
<td>0.68</td>
<td>0.48</td>
</tr>
<tr>
<td>A1d</td>
<td>0.89</td>
<td>0.32</td>
</tr>
<tr>
<td>A1e</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>A2a</td>
<td>0.79</td>
<td>0.42</td>
</tr>
<tr>
<td>A2b</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>A2c</td>
<td>0.37</td>
<td>0.50</td>
</tr>
<tr>
<td>A3a</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>Difficulty</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>A4a</td>
<td>0.47</td>
<td>0.51</td>
</tr>
<tr>
<td>A5a</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>A5b</td>
<td>0.58</td>
<td>0.51</td>
</tr>
<tr>
<td>A5c</td>
<td>0.89</td>
<td>0.32</td>
</tr>
<tr>
<td>A5d</td>
<td>0.84</td>
<td>0.37</td>
</tr>
<tr>
<td>A5e</td>
<td>0.79</td>
<td>0.42</td>
</tr>
<tr>
<td>A6a</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>A6b</td>
<td>0.95</td>
<td>0.23</td>
</tr>
<tr>
<td>A6c</td>
<td>0.84</td>
<td>0.37</td>
</tr>
<tr>
<td>A6d</td>
<td>0.32</td>
<td>0.48</td>
</tr>
<tr>
<td>B1a</td>
<td>0.84</td>
<td>0.37</td>
</tr>
<tr>
<td>B1b</td>
<td>0.79</td>
<td>0.42</td>
</tr>
<tr>
<td>B1c</td>
<td>0.95</td>
<td>0.23</td>
</tr>
<tr>
<td>B1d</td>
<td>0.89</td>
<td>0.32</td>
</tr>
<tr>
<td>B2a</td>
<td>0.68</td>
<td>0.48</td>
</tr>
<tr>
<td>B2b</td>
<td>0.84</td>
<td>0.37</td>
</tr>
<tr>
<td>B3a</td>
<td>0.84</td>
<td>0.37</td>
</tr>
<tr>
<td>Checklist</td>
<td>20.37</td>
<td>2.69</td>
</tr>
<tr>
<td>GAS</td>
<td>2.53</td>
<td>0.51</td>
</tr>
<tr>
<td>Total</td>
<td>22.89</td>
<td>2.94</td>
</tr>
</tbody>
</table>

Table 14.8 and Figure 14.3 show the spread of mean item scores. Only five of the twenty-six checklist items were completed by half the participants or fewer (mean item value ≤ 0.50) and seventeen of the items were completed by more than 80% of participants (mean item value > 0.80).
Table 14.8: Spread of item mean item values for Simulation Scenario 1

<table>
<thead>
<tr>
<th>Mean Item Value</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.31-0.40</td>
<td>3</td>
</tr>
<tr>
<td>0.41-0.50</td>
<td>1</td>
</tr>
<tr>
<td>0.51-0.60</td>
<td>2</td>
</tr>
<tr>
<td>0.61-0.70</td>
<td>2</td>
</tr>
<tr>
<td>0.71-0.80</td>
<td>3</td>
</tr>
<tr>
<td>0.81-0.90</td>
<td>9</td>
</tr>
<tr>
<td>0.91-1.0</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 14.3: Spread of mean item values for the Simulation Scenario 1

Table 14.9 shows the topics and categories of learning outcomes covered, sorted by mean item value. For all learning outcome categories that were assessed by more than one item, there was a range of mean item values, thereby allowing differentiation of candidates’ performance in each of these learning outcome categories. There was a very even spread of mean item values for the history component, and a reasonable spread for other aspects of patient assessment. However, mean item values were mostly higher for the patient management learning outcome categories.
Table 14.9: Topics and categories of learning outcomes covered for Simulation Scenario 1, sorted by mean item value

Note: * Item B1d falls under two different categories of learning outcomes

<table>
<thead>
<tr>
<th>Learning Outcome Category</th>
<th>MEAN ITEM VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.31-0.40</td>
</tr>
<tr>
<td></td>
<td>0.41-0.50</td>
</tr>
<tr>
<td></td>
<td>0.51-0.60</td>
</tr>
<tr>
<td></td>
<td>0.61-0.70</td>
</tr>
<tr>
<td></td>
<td>0.71-0.80</td>
</tr>
<tr>
<td></td>
<td>0.81-0.90</td>
</tr>
<tr>
<td></td>
<td>0.91-1.0</td>
</tr>
</tbody>
</table>

**Item Numbers**

**A. Patient Assessment**

1. History
   - A2c
   - A3a
   - A4a
   - A2b
   - A1c
   - A2a
   - A1a
   - A1d
   - A1b
   - A1e

2. Examination
   - A5b
   - A5e
   - A5c
   - A5d
   - A5a

3. Investigation
   - A6d

4. Diagnosis
   - A6b

**B. Patient management**

1. Management
   - B1b
   - B1a
   - B1c
   - B1d*

2. Patient progress
   - B2a
   - B1d*
   - B2b

3. Prioritisation
   - B3a

There were four items on which MD3 participants performed better than HMO2s, eleven items on which there was no difference in mean performance between the two groups, and eleven items on which HMO2 participants performed better than MD3s. A Mann-Whitney test comparing performance of MD3 and HMO2 participants on each item of the Simulation Scenario 2 assessment tool demonstrated that only two items (items A1c and A4a) had statistically significant differences in results for MD3 and HMO2 groups (p < 0.05), with HMO2s performing better than MD3s for on item A4a, but MD3s performing better than HMO2s on item A1c. Item A1c required the participant to enquire as to the character of the patient’s chest pain (to help distinguish from other causes) and item A4a required the participant to ask about previous history of ischaemic heart disease (for the purposes of risk stratification). Results for these items are shown in Table 14.10.
Table 14.10: Results for MD3 and HMO2 participants for Simulation Scenario 1 items A1c and A4a

<table>
<thead>
<tr>
<th>Item</th>
<th>Class</th>
<th>n</th>
<th>Mean</th>
<th>Mann-Whitney p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1c</td>
<td>MD3</td>
<td>4</td>
<td>1.00</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>A4a</td>
<td>MD3</td>
<td>4</td>
<td>0.25</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

A Kruskal-Wallis test comparing performance across all classifications (MD2 to HMO2) demonstrated that two items (items A1c and A5c) had statistically significant differences in results across the classifications. Item A1c required the participant to enquire as to the character of the patient’s chest pain, and item A5c required the participant to auscultate the heart. Results for these items are shown in Table 14.11.

Table 14.11: Results for participants of all classifications for Simulation Scenario 1 items A1c and A5d

Note: * MD2 score is for an individual participant

<table>
<thead>
<tr>
<th>Item</th>
<th>Class</th>
<th>n</th>
<th>Mean</th>
<th>Kruskal-Wallis p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1c</td>
<td>MD2</td>
<td>1</td>
<td>0.00*</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>4</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD4</td>
<td>3</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>6</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>19</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>A5c</td>
<td>MD2</td>
<td>1</td>
<td>0.00*</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>4</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MD4</td>
<td>3</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>6</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>5</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>19</td>
<td>0.89</td>
<td></td>
</tr>
</tbody>
</table>
For each item that demonstrated a significant difference across all classification, a pairwise comparison was used to determine the specific groups with significantly different results.

Kruskal-Wallis test comparison pairwise across all classifications, showed no significant difference between any two groups for Scenario 1 Item A1c (all adjusted significance values were > 0.05).

Kruskal-Wallis test comparison pairwise across all classifications, showed significant differences (p < 0.05) on Scenario 1 Item A5c for the following pairs of groups:

- MD2 and MD3: adjusted significance 0.046
- MD2 and HMO2: adjusted significance 0.038
ii. Quantitative Component: Simulation Scenario 2

Table 14.12 shows the mean item value and standard deviation for each item of Simulation Scenario 1. In the context of a simulation assessment checklist, mean item values are a measure of the proportion of examinees satisfactorily executing a particular step, rather than item ‘difficulty’.

Table 14.12: Mean item value and standard deviation for each item of Simulation Scenario 2

<table>
<thead>
<tr>
<th>Item/Component</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1a</td>
<td>0.67</td>
<td>0.49</td>
</tr>
<tr>
<td>A1b</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>A2a</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td>A2b</td>
<td>0.94</td>
<td>0.24</td>
</tr>
<tr>
<td>A2c</td>
<td>0.22</td>
<td>0.43</td>
</tr>
<tr>
<td>A2d</td>
<td>0.33</td>
<td>0.49</td>
</tr>
<tr>
<td>A2e</td>
<td>0.22</td>
<td>0.43</td>
</tr>
<tr>
<td>A2f</td>
<td>0.11</td>
<td>0.32</td>
</tr>
<tr>
<td>A2g</td>
<td>0.06</td>
<td>0.24</td>
</tr>
<tr>
<td>A3a</td>
<td>0.94</td>
<td>0.24</td>
</tr>
<tr>
<td>A3b</td>
<td>0.56</td>
<td>0.51</td>
</tr>
<tr>
<td>A3c</td>
<td>0.72</td>
<td>0.46</td>
</tr>
<tr>
<td>A3d</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>A3e</td>
<td>0.83</td>
<td>0.38</td>
</tr>
<tr>
<td>A4a</td>
<td>0.72</td>
<td>0.46</td>
</tr>
<tr>
<td>A4b</td>
<td>0.78</td>
<td>0.43</td>
</tr>
<tr>
<td>A4c</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>A4d</td>
<td>0.28</td>
<td>0.46</td>
</tr>
<tr>
<td>A4e</td>
<td>0.06</td>
<td>0.24</td>
</tr>
<tr>
<td>B1a</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>B1b</td>
<td>0.78</td>
<td>0.43</td>
</tr>
<tr>
<td>B1c</td>
<td>0.83</td>
<td>0.38</td>
</tr>
<tr>
<td>B1d</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>B1e</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td>B2a</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td>B2b</td>
<td>0.78</td>
<td>0.43</td>
</tr>
<tr>
<td>B3a</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>Checklist</td>
<td>14.67</td>
<td>2.11</td>
</tr>
<tr>
<td>GAS</td>
<td>1.39</td>
<td>0.70</td>
</tr>
<tr>
<td>Total</td>
<td>16.06</td>
<td>2.67</td>
</tr>
</tbody>
</table>
Table 14.13 and Figure 14.4 show the spread of mean item scores. Twelve of the twenty-seven checklist items were completed by half the participants or fewer (mean item value $\leq 0.50$) and only five of the items were completed by more than 80% of participants (mean item value $> 0.80$). This result is in contrast to the distribution of mean item values for Scenario 1, where there were relatively few items which were completed by half the participants or fewer, and a large number of items which were completed by more than 80% of participants.

<table>
<thead>
<tr>
<th>Mean Item Value</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00-0.10</td>
<td>3</td>
</tr>
<tr>
<td>0.11-0.20</td>
<td>1</td>
</tr>
<tr>
<td>0.21-0.30</td>
<td>3</td>
</tr>
<tr>
<td>0.31-0.40</td>
<td>1</td>
</tr>
<tr>
<td>0.41-0.50</td>
<td>4</td>
</tr>
<tr>
<td>0.51-0.60</td>
<td>1</td>
</tr>
<tr>
<td>0.61-0.70</td>
<td>4</td>
</tr>
<tr>
<td>0.71-0.80</td>
<td>5</td>
</tr>
<tr>
<td>0.81-0.90</td>
<td>2</td>
</tr>
<tr>
<td>0.91-1.0</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 14.4: Spread of mean item values for the Simulation Scenario 2
Table 14.14 shows the topics and categories of learning outcomes covered, sorted by mean item value. For all learning outcome categories that were assessed by more than one item, there was a range of mean item values, thereby allowing differentiation of candidates’ performance in each of these learning outcome categories. As for Simulation Scenario 1, there was a tendency for higher mean item values for those items that assessed patient management. This suggests that participants were more comprehensive in management, compared to assessment.

<table>
<thead>
<tr>
<th>Learning Outcome Category</th>
<th>MEAN ITEM VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤ 0.10</td>
</tr>
<tr>
<td>A. Patient Assessment</td>
<td></td>
</tr>
<tr>
<td>1. History</td>
<td>A1b, A2g</td>
</tr>
<tr>
<td>2. Examination</td>
<td></td>
</tr>
<tr>
<td>3. Investigation</td>
<td>A4e, A4d, A4c</td>
</tr>
<tr>
<td>4. Diagnosis</td>
<td></td>
</tr>
<tr>
<td>B. Patient management</td>
<td></td>
</tr>
<tr>
<td>1. Management</td>
<td></td>
</tr>
<tr>
<td>2. Patien progress</td>
<td></td>
</tr>
<tr>
<td>3. Prioritisation</td>
<td></td>
</tr>
</tbody>
</table>
There were nine items on which MD2 participants performed better than HMO2s, two items on which there was no difference in mean performance (rank) between the two groups, and sixteen items on which HMO2 participants performed better than MD2s. A Mann-Whitney test comparing performance of MD2 and HMO2 participants on each item of the Simulation Scenario 2 assessment tool demonstrated that only two items (items A4a and A4c) had statistically significant differences in results for MD2 and HMO2 groups (p < 0.05), with HMO2s performing better than MD2s for both of these items. Items A4a and A4c required the participant to request an ECG and chest x-ray, respectively. Results for these items are shown in Table 14.15.

### Table 14.15: Results for MD2 and HMO2 participants for Simulation Scenario 2 items A4a and A4c

<table>
<thead>
<tr>
<th>Item</th>
<th>Class</th>
<th>N</th>
<th>Mean</th>
<th>Mann-Whitney p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4a</td>
<td>MD2</td>
<td>7</td>
<td>0.29</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>A4c</td>
<td>MD2</td>
<td>7</td>
<td>0.00</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>0.83</td>
<td></td>
</tr>
</tbody>
</table>

A Kruskal-Wallis test comparing performance across all classifications (MD2 to HMO2) demonstrated that four items (items A4a, A4b, A4c, A4d) had statistically significant differences in results across the classifications. Item A4a required the participant to request an ECG, item A4b required recognition of atrial fibrillation, item A4c required the participant to request a chest x-ray and item A4d requesting of blood tests. Results for these items are shown in Table 14.16.
Table 14.16: Results for participants of all classifications for Simulation Scenario 2 items A4a, A4b, A4c and A4d

Note: No MD4s were allocated to Simulation Scenario 2

<table>
<thead>
<tr>
<th>Item</th>
<th>Class</th>
<th>n</th>
<th>Mean</th>
<th>Kruskal-Wallis p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4a</td>
<td>MD2</td>
<td>7</td>
<td>0.29</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>2</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>3</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>A4b</td>
<td>MD2</td>
<td>7</td>
<td>0.71</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>2</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>3</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>A4c</td>
<td>MD2</td>
<td>7</td>
<td>0.00</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>2</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>3</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>A4d</td>
<td>MD2</td>
<td>7</td>
<td>0.00</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>MD3</td>
<td>2</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intern</td>
<td>3</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HMO2</td>
<td>6</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>18</td>
<td>0.28</td>
<td></td>
</tr>
</tbody>
</table>

For each item that demonstrated a significant difference across all classifications, a pairwise comparison was used to determine the specific groups with significantly different results.
Kruskal-Wallis test comparison pairwise across all classifications, showed no significant difference between any two groups for Scenario 2 Item A4a (all adjusted significance values were > 0.05).

Kruskal-Wallis test comparison pairwise across all classifications, showed significant differences (p < 0.05) on Scenario 2 Item A4b for the following pair of groups:

- MD3 and HMO2: adjusted significance 0.042

Kruskal-Wallis test comparison pairwise across all classifications, showed significant differences (p < 0.05) on Scenario 2 Item A4c for the following pairs of groups:

- MD2 and Intern: adjusted significance 0.049
- MD2 and HMO2: adjusted significance 0.036

Kruskal-Wallis test comparison pairwise across all classifications, showed significant differences (p < 0.05) on Scenario 2 Item A4d for the following pair of groups:

- MD2 and Intern: adjusted significance 0.017
iii. Qualitative Component

Figure 14.5, below, shows the findings of the qualitative analysis of the feedback for the simulation scenarios (based on the instructions for debrief supplied to the examiners in the simulation scenario document). The four sections of the figure show the domains of feedback that are assessed in the evaluation of formative validity: form, type, quantity and quality. Within each of these domains are descriptive findings of the qualitative analysis, which show the characteristics of the feedback provided to participants during the simulation debrief.

![Figure 14.5: Analysis of the simulation scenario feedback](image)

The form of the feedback provided to participants was an immediate, verbal, structured debrief by the assessors. Individual’s marks were not disclosed to them, nor was any comparison made with the other participants. The type of feedback provided was individualised bi-directional/discussion-based feedback, which required the participants to reflect on their performance, before being offered both positive and negative...
feedback by the assessors. Where possible and indicated, individual teaching on very specific areas of weakness was provided.

The quantity of feedback was provided is this time described in terms of time and number of feedback/teaching points (as compared to the description of the amount of written material provided through the MCQ test feedback document). The debrief was conducted over a 5 minute period, during which assessors were instructed to address four specific points of feedback and provide teaching on at least two points. The quality of the feedback is described in terms of the quality of the assessors providing feedback and the basis of the feedback provided. As shown in Figure 14.5, at least one assessor for each simulation scenario had experience in cardiology, and feedback provided was based on clinical experience, local clinical policy/practice, and guidelines.

The implications of the findings regarding the simulation assessment feedback and formative validity are discussed in Chapter 20.
CHAPTER 15: INTEROBSERVER RELIABILITY

Table 15.1, below, shows the percentage agreement for each combination of assessors, for each component of the simulation assessment tools and for each simulation scenario, as well as both simulation scenarios combined. The percentage agreement represents the proportion of participant scores that correlate exactly between the two assessors (i.e. both assessors gave exactly the same score). The percentage agreement between assessors varied significantly by assessment tool component and assessor combination, but was generally highest (scores were most concordant) for the combination of assessors 1 and 5.

Assessor 4 was not included in the analysis, since this observer only assessed two participants. Similarly, interobserver reliability was not calculated for the combination of Assessor 1 and Assessor 2 for individual scenarios, due to small numbers (as discussed in Chapter 9).

Table 15.1: Percentage agreement for each combination of assessors

<table>
<thead>
<tr>
<th>Assessors</th>
<th>Simulation Scenario</th>
<th>n</th>
<th>Overall</th>
<th>Checklist</th>
<th>GAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>Combined</td>
<td>5</td>
<td>20</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>Combined</td>
<td>15</td>
<td>26.7</td>
<td>33.3</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Scenario 1</td>
<td>7</td>
<td>28.6</td>
<td>28.6</td>
<td>42.9</td>
</tr>
<tr>
<td></td>
<td>Scenario 2</td>
<td>8</td>
<td>25</td>
<td>37.5</td>
<td>37.5</td>
</tr>
<tr>
<td>1 &amp; 5</td>
<td>Combined</td>
<td>12</td>
<td>50</td>
<td>41.7</td>
<td>58.3</td>
</tr>
<tr>
<td></td>
<td>Scenario 1</td>
<td>6</td>
<td>50</td>
<td>50</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>Scenario 2</td>
<td>6</td>
<td>50</td>
<td>33.3</td>
<td>50</td>
</tr>
</tbody>
</table>
Table 15.2, below, shows the Spearman correlation for each combination of assessors for each component of the simulation assessment tools and for each simulation scenario. Spearman correlations could not be performed for both scenarios combined since Spearman correlations use the ranking of scores, and scores from the two different assessments could not be ranked together, since their distribution was different. It is important to note that where the correlation is 1.00 (i.e. perfect correlation), the p-value cannot be calculated, but is significant, since the highest possible correlation has been achieved.

Table 15.2 demonstrates that the Spearman’s correlation varied in magnitude for the different assessment tool components and different combinations of assessors. The correlation was strong or better ($\rho > 0.60$), and significant ($p < 0.05$) for:
- assessors 1 and 3 for the overall score
- assessors 1 and 3 for the checklist component for Scenario 1
- assessors 1 and 3 for the checklist component for Scenario 2

The correlation was perfect ($\rho = 1.00$) for the combination of assessors 1 and 5 for the overall and checklist scores. No other correlations were statistically significant.

Table 15.2: Spearman correlation for each combination of assessors
Note: * where the correlation is 1.00, the p-value cannot be calculated

<table>
<thead>
<tr>
<th>Assessors</th>
<th>Simulation Scenario</th>
<th>n</th>
<th>Overall</th>
<th>Checklist</th>
<th>GAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>spearman’s correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>Scenario 1</td>
<td>7</td>
<td>$0.97$</td>
<td>0.000</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Scenario 2</td>
<td>8</td>
<td>$0.66$</td>
<td>0.076</td>
<td>0.76</td>
</tr>
<tr>
<td>1 &amp; 5</td>
<td>Scenario 1</td>
<td>6</td>
<td>$1.00$</td>
<td>*</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Scenario 2</td>
<td>6</td>
<td>$0.79$</td>
<td>0.063</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Table 15.3, below, shows Cohen’s kappa for each combination of assessors for each component of the simulation assessment tools and for each simulation scenario, as well as both simulation scenarios combined. Cohen’s kappa was of low magnitude (poor correlation, $\kappa < 0.40$) for most components.

Cohen’s kappa indicated a correlation that was fair or better ($\kappa > 0.40$), and statistically significant ($p < 0.05$), for:

- assessors 1 and 2 for the checklist component for both scenarios combined
- assessors 1 and 5 for the overall score for both scenarios combined
- assessors 1 and 5 for the overall score for Scenario 1
- assessors 1 and 5 for the checklist component for Scenario 1

Table 15.3: Cohen’s kappa for each combination of assessors

<table>
<thead>
<tr>
<th>Assessors</th>
<th>Simulation Scenario</th>
<th>n</th>
<th>Overall Cohen’s Kappa</th>
<th>Checklist Cohen’s Kappa</th>
<th>GAS Cohen’s Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\kappa$</td>
<td>p-value</td>
<td>$\kappa$</td>
</tr>
<tr>
<td>1 &amp; 2</td>
<td>Combined</td>
<td>5</td>
<td>0.17</td>
<td>0.025</td>
<td>0.55</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>Combined</td>
<td>15</td>
<td>0.20</td>
<td>0.009</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Scenario 1</td>
<td>7</td>
<td>0.19</td>
<td>0.15</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Scenario 2</td>
<td>8</td>
<td>0.13</td>
<td>0.33</td>
<td>0.23</td>
</tr>
<tr>
<td>1 &amp; 5</td>
<td>Combined</td>
<td>12</td>
<td>0.46</td>
<td>0.000</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Scenario 1</td>
<td>6</td>
<td>0.46</td>
<td>0.000</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>Scenario 2</td>
<td>6</td>
<td>0.36</td>
<td>0.071</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Table 15.4 shows the percentage agreement for each combination of assessors, for each item on each simulation scenario. Scenario 1 had ten items (out of 26) with overall percentage agreement of 100% and Scenario 2 had 13 items (out of 27) with overall percentage agreement of 100%.

Table 15.4: Percentage agreement for each item
Notes: Scenario 1 item 3a has a maximum attainable mark of 2; there were three cases for which there was only a single assessor

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Assessors</th>
<th>Percentage agreement</th>
<th>Scenario 2</th>
<th>Assessors</th>
<th>Percentage agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 &amp; 3</td>
<td>1 &amp; 5</td>
<td></td>
<td>1 &amp; 3</td>
<td>1 &amp; 5</td>
</tr>
<tr>
<td>Item</td>
<td>n</td>
<td>Overall</td>
<td>n</td>
<td>Overall</td>
<td></td>
</tr>
<tr>
<td>A1a</td>
<td>71.4</td>
<td>83.3</td>
<td>100.0</td>
<td>62.5</td>
<td>0.0</td>
</tr>
<tr>
<td>A1b</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A1c</td>
<td>85.7</td>
<td>83.3</td>
<td>88.2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A1d</td>
<td>100.0</td>
<td>83.3</td>
<td>94.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A1e</td>
<td>100.0</td>
<td>100.0</td>
<td>94.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A2a</td>
<td>85.7</td>
<td>83.3</td>
<td>82.4</td>
<td>82.4</td>
<td>100.0</td>
</tr>
<tr>
<td>A2b</td>
<td>100.0</td>
<td>100.0</td>
<td>94.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A2c</td>
<td>71.4</td>
<td>100.0</td>
<td>88.2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A3a</td>
<td>85.7</td>
<td>100.0</td>
<td>94.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A4</td>
<td>100.0</td>
<td>66.7</td>
<td>76.5</td>
<td>87.5</td>
<td>100.0</td>
</tr>
<tr>
<td>A5a</td>
<td>85.7</td>
<td>100.0</td>
<td>94.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A5b</td>
<td>100.0</td>
<td>66.7</td>
<td>88.2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A5c</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A5d</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A5e</td>
<td>100.0</td>
<td>83.3</td>
<td>88.2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A6a</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>87.5</td>
<td>83.3</td>
</tr>
<tr>
<td>A6b</td>
<td>85.7</td>
<td>100.0</td>
<td>94.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>A6c</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>87.5</td>
<td>100.0</td>
</tr>
<tr>
<td>A6d</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>87.5</td>
<td>100.0</td>
</tr>
<tr>
<td>B1a</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>87.5</td>
<td>100.0</td>
</tr>
<tr>
<td>B1b</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>B1c</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>B1d</td>
<td>85.7</td>
<td>100.0</td>
<td>88.2</td>
<td>87.5</td>
<td>66.7</td>
</tr>
<tr>
<td>B2a</td>
<td>85.7</td>
<td>100.0</td>
<td>94.1</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>B2b</td>
<td>57.1</td>
<td>83.3</td>
<td>76.5</td>
<td>75.0</td>
<td>66.7</td>
</tr>
<tr>
<td>B3a</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>50.0</td>
<td>66.7</td>
</tr>
</tbody>
</table>

| Overall    | 87.5      | 100.0                | 94.1       | 87.5      | 100.0                |
CHAPTER 16: QUALITATIVE ANALYSIS OF THE RESEARCH PROCESS

Results of the qualitative analysis are presented in diagrammatic form, showing the taxonomies that describe the phenomena of *participant recruitment and logistics* and *simulation assessment execution*.

Figure 16.1: Diagrammatic representation of the domains and dimensions derived from qualitative data on the topic of *participant recruitment and logistics*.
Figure 16.1 demonstrates the taxonomy describing the phenomenon of *participant recruitment and logistics*, with the domains and dimensions derived from the qualitative analysis of the topic. The taxonomy is shown in the form of a Venn diagram, in order to depict areas of overlap and relationships. The major domains, derived from the codes assigned to the qualitative data, are each represented by a different coloured circle and labelled in bold, capitalised lettering. The dimensions within each domain, derived from the subcodes assigned to the data, are labelled in lower case lettering, and their position demonstrates their relationship to the domains. For example, if a dimension relates to more than one domain, then it is located in the area of overlap between the related domains.

Figure 16.1 shows that there are four conceptual domains, which contribute to and influence *participant recruitment*. These are: time, resources, interest, and communication. The descriptions of each of the dimensions, as they relate to their specific domains, are listed below. An asterisk denotes overlap with another domain:

1. Time:  
   A. Timeframe- the timeframe over which the research was undertaken  
   B. Sessions- the timing/scheduling of the data collection sessions  
   C. Resource availability*- referring to time as a resource, and the limited availability of participants and assessors  
   D. Logistics*- the practicalities of timing/timetabling and the constraints conferred by logistical requirements  
   E. Competing priorities*- in this domain, relates to the multiple priorities competing for participants’ time
2. Resources:

A. Simulation lab - the simulation lab as a resource required in order to run the data collection sessions

B. Assessors - the assessors as resources required to run the data collection sessions

C. Actors - the actors as resources required to run the data collection sessions

D. Resource availability* - the availability of various resources required to run the data collection sessions; in this domain referring to physical and human resources

E. Logistics* - the logistics of co-ordinating the availability of physical and human resources

3. Interest:

A. Opportunity - potential participants’ perception of the simulation session as an opportunity (for learning, practice, experience, etc.)

B. Learning - potential participants’ perception of the value of the simulation session for learning

C. Word of mouth* - the impact of informal advertising (positive or negative) on potential participants’ interest in participating

D. Competing priorities* - the competing interests for potential participants’ interests (e.g. other research projects, other learning opportunities)
4. Communication:

A. Recruitment material- the communication to potential participants via the recruitment material

B. Word of mouth*- the communication received by potential participants via informal advertising

C. Competing priorities*- the effect of competing communications being received by potential participants

Figure 16.2, below, demonstrates the domains and dimensions on the topic of simulation assessment execution. The figure is in the form of a mind map, which shows how each domain represents a component of simulation, and within each domain there are dimensions, which are coloured according to their relationship to the broader domain.

![Figure 16.2: Map of domains and dimensions derived from qualitative analysis on the topic of simulation assessment execution](image-url)
Figure 16.2 shows that there are five conceptual domains, which contribute to and influence simulation assessment execution. These are: design, execution, expectations and performance, feedback, and assessment. The descriptions of each of the dimensions, as they relate to their specific domains, are listed below.

1. Design:
   A. Difficulty- the difficulty of the simulation assessments
   B. Fidelity- the fidelity with which the simulation scenarios are designed to be presented during the simulation (aspects inherent to design)
   C. Consistency- the level of consistency in representation of the clinical scenarios that is dependent on simulation scenario design (as opposed to execution)

2. Execution:
   A. Fidelity- the fidelity with which the simulation scenarios were executed (aspects dependent on execution rather than design)
   B. Reproducibility/consistency- the consistency with which the simulation scenarios were portrayed, the practical reproducibility of the scenarios
   C. Actors- the impact that the actors had on the execution of the simulation scenarios
   D. Unexpected outcome- an unexpected course taken by a participant in completing the simulation scenario

3. Expectations and performance:
   A. Participants’ expectations- participants’ expectations of what the simulations would be like
   B. Performance- participants’ performance on the simulations
4. Feedback:  
   A. Participants- participants’ feedback regarding the simulation scenarios/assessments
   
   B. Assessors- assessors’ feedback regarding the simulation scenarios/assessments

5. Assessment:  
   A. Assessment tool- the simulation assessments as assessment tools
   
   B. Formative assessment- the formative aspects of the simulation assessments

Each of the domains and dimensions identified in Figures 16.1 and 16.2 are discussed in detail in the discussion section of the thesis (Chapter 22).
SECTION IV: DISCUSSION

CHAPTER 17: PARTICIPANT CHARACTERISTICS AND ASSESSMENT

RESULTS

This chapter will address Research Question 2- “What is the relationship between performance on simulation-based and MCQ assessments of cardiac emergency management?”

A. Participant Characteristics

i. Participant Demographics

The majority of participants recruited for this study were female (63.2%). Data regarding gender proportions was available for all eligible participant cohorts except for HMO2. Overall, for the medical student and intern cohorts combined, 54.9% of potential participants were female. Therefore, the female majority observed can be attributed to overall female majority in the eligible cohorts.

The mean participant age of 25.8 years, and range of ages (21-33) reflects that all medical student participants were drawn from the University of Melbourne graduate-entry medicine course, and 76.2% of the JMOs had completed a graduate-entry medical course. This may not be consistent with other cohorts in Australia, particularly in those states that offer a different mix of undergraduate and graduate-entry medical training (Victoria offers predominantly graduate-entry medical degrees). Despite an increasing number of graduate-entry medical courses available in Australia, the median age of graduates has remained steady at 25 years old over the period 2010-2014, though the age range has increased over the same time. (99)
It is important to note that participants in this study included medical students in their last three years of study as well as junior doctors, so the age data should not be compared directly with data that relates only to graduates. Median age of those participants in their final year of medical school was not calculated as there were only three MD4 participants.

In regards to the classifications of participants (based on experience and, for HMO2s, stream), there were proportionally more intern and HMO2 participants, compared to medical students. There were very few MD4 and HMO2 surgical participants. Conversely, most of the HMO2 participants (7 of 11, 63.6%) were from the medical stream. It is beyond the scope of this research to draw conclusions as to the reasons why certain groups were more successfully recruited.

All medical students at the Northern Clinical School are students of the University of Melbourne MD course, with the exception of occasional elective students from other institutions, who were not eligible to participate in this research. As such, there was no variation in the medical degree of the students who participated.

Additional demographic data was collected as part of the participant questionnaire, including participants’ previous experience (working in a hospital, both in medical and other capacities; experience in relevant rotations). This was initially intended to allow for further subgroup analysis if larger participant numbers were recruited, however, low participant numbers meant that this was not feasible.
ii. Simulation Scenario Groups

There were no statistically significant differences in the demographics of the groups allocated to the two simulation scenarios. There were some differences in the number of participants of each classification who were allocated to each scenario; again, this did not reach statistical significance (p = 0.055), but proved to be a limitation for some statistical analyses. Allocation of MD2 and MD4 students to the two simulation scenarios was particularly uneven, due to the allocation method used. Of the eight MD2 students who completed the simulation assessments, only one was allocated to Scenario 1 and seven were allocated to Scenario 2. This meant that there was no group data available for MD2 students’ performance on Scenario 1 (only an individual participant’s performance data). All three of the MD4 students who participated were allocated to Scenario 1, so there is no data available on the performance of the MD4 group on Scenario 2.

B. Multiple Choice Question Test

Prior to undertaking further statistical analysis, it was pertinent to determine the distribution of the results of the MCQ test. Many statistical analyses assume normal distribution of the data and may not be able to be accurately applied or interpreted if fulfilment of this assumption is not proven.(83, 84)

Both the histogram and the Shapiro-Wilk test demonstrated that the results of the MCQ test for the entire cohort were approximately normally distributed. An approximately normal distribution of results suggests that the MCQ test was able to differentiate participants throughout the spectrum of achievement and, in particular, differentiate amongst the best performing and poorest performing participants.(83) This is an important concept in norm referencing, where a student’s performance is measured by
his/her relative standing within the cohort. (11, 100) This is in contrast to criterion-referencing, where the relative standing of a student within the cohort is not important, and achievement of pre-determined criteria is the only measure of achievement. (11, 100)

Identification of poorer-performing examination candidates is important in both formative and summative assessments. In a formative assessment, identifying those learners who may be struggling to grasp the content being taught is beneficial to both the learner and the educator. If the learner is able to identify that he/she is performing poorly and receive feedback on areas and methods for improvement, this is more likely to provide motivation and guidance for further learning. (100) Similarly, it is important for educators to identify those learners who may require additional support, and subsequently, methods to provide this support. (11) In summative assessment, especially if norm referencing is used, it is important to be able to clearly determine which candidates’ achievement is inadequate to pass the assessment. (11)

C. Simulation Assessments

The two components of the simulation assessment score— the checklist component and the GAS— were examined individually, and combined as a total score. Both the histogram and the Shapiro-Wilk test demonstrated that the results for the checklist components and total scores for each simulation scenario for the entire cohort were approximately normally distributed. Results for the GAS for each simulation scenario were not normally distributed (Shapiro-Wilk significance < 0.05). This may attributable to the very small scale of the GAS, which has minimum achievable mark of 0, and maximum achievable mark of 3.
There was a large difference in mean scores achieved for the two simulation scenarios (Table 11.3). The mean total score for Simulation Scenario 1 was considerably higher than for Simulation Scenario 2. However, the magnitude of the range of marks was similar for both scenarios (Simulation Scenario 1 range = 9, Simulation Scenario 2 range = 10). This suggests that both simulation scenarios were equally good at differentiating between candidates, but Simulation Scenario 2 was more difficult than Simulation Scenario 1.

Similarly, the mean GAS was considerably higher for Simulation Scenario 1, compared to Simulation Scenario 2. However, unlike the total score, the GAS score range was much smaller for Scenario 1 than for Scenario 2. Given GAS assesses overall proficiency/competence, the findings from the GAS scores suggest, firstly, that participants were more proficient at managing an AMI scenario than an APO scenario, and secondly, that there was greater variation in participants’ overall proficiency at APO management. Again, this reinforces that Simulation Scenario 2 was more difficult than Simulation Scenario 1.

Although the range of marks for the total score for the two simulation scenarios was similar, the range (9-10) was significantly smaller than the possible range of marks (0-30). Given that there was very little subjectivity in the checklist component of the score, which makes up 27 of the 30 marks, the small range of results is likely to be a true reflection of a relatively homogenous performance across the cohorts. Another possible explanation for a small range would be a central tendency effect (examiners not using the entire range of the marking scale), but this could only apply to the more subjective GAS component. (86)
In regards to correlation between the two components of the simulation assessment, there was statistically significant, and relatively strong, correlation between the checklist score and GAS for Simulation Scenario 2, but not for Simulation Scenario 1. The correlation between checklist score and GAS for Simulation Scenario 2 was strong and was statistically significant. Conversely, the correlation between checklist score and GAS for Simulation Scenario 1 was low and not statistically significant. The reason for such large difference in the strength of checklist/GAS correlation for the two simulation scenarios is unclear, but this may relate to the magnitude of the range of scores for the two scenarios. As previously mentioned, the magnitude of the range of total scores was very similar for both scenarios. This was also the case for the checklist scores. However, the range for the GAS component was much smaller for Simulation Scenario 1. The smaller magnitude of the range of GAS scores for Simulation Scenario 1 means that the much larger range of checklist marks has been condensed to a range of only two GAS marks, and correlation over such a small range is unlikely.

This raises the question of whether the GAS alone is able to capture the complexity of candidate performance that is usually assessed in the form of a checklist, and perhaps stratify performance on a simpler scale. While criteria are clearly stated for the assignment of the GAS score (Appendix 5: Simulation Scenarios), there is more subjectivity in determination of this score, compared to the checklist scoring, which requires only binary determination of whether or not the checklist items were completed (with the exception of Item A3a in Simulation Scenario 1, which is marked 0-2). A score with such subjectivity, and small range of available marks, is not desirable for formal assessment, where performance across multiple criteria should be used to assess candidates, in order to optimise reliability.
Despite intending to avoid consequential marking in the design of the simulation assessment tool, in particular by allocating more marks for general patient assessment, and allocating a mark for seeking senior assistance, it is clear that some of the marking criteria in the simulation checklists did rely on completion of an earlier criterion. For example, the marks allocated in each simulation for recognition of ECG abnormalities did assume that the candidates would first request this investigation. Therefore, those candidates who did not request an ECG could not, in theory, have fulfilled the criterion for correct interpretation of the ECG. It should be noted, however, that for Scenario 2, some candidates did state that the patient was in atrial fibrillation based on the trace displayed on the monitor, without requesting an ECG, and therefore obtained the mark allocated for recognition of AF without obtaining the mark for requesting an ECG. These candidates were subsequently advised during debrief that while their diagnosis was correct, atrial fibrillation should properly be diagnosed on a 12 lead ECG.

Marks allocated for many of the management components of the checklist also relied heavily on candidates recognising the patient’s pathology at an earlier stage, based on completion of earlier criteria. However, again, if participants failed to recognise the pathology but instituted correct treatment (as occurred in one case), they were not penalised in terms of checklist marks (but should have scored lower on the GAS component). While it may not be possible to fully circumvent the issue of consequential marking in clinical simulations, it is important to recognise the impact this has on participant scores.
D. Correlation Between MCQ test and Simulation Assessments

The moderate correlation between the two assessment methods (MCQ and simulation) suggests that they are assessing similar constructs. (101) The MCQ and simulation assessment tools were designed to assess cardiac emergency management, specifically AMI and APO management. However, the MCQ test assesses both AMI and APO management, whereas these topics are assessed by separate simulation assessments. Therefore, it is expected that only half the MCQ test (the fifteen items related to AMI) would correlate to Simulation Scenario 1, and the other half of the test (the fifteen items related to APO) would correlate to Scenario 2. Bearing this in mind, the correlation values obtained are actually higher than would be expected.

While the constructs are similar, in terms of the knowledge, topics and concepts assessed, as previously discussed, the two assessment methods assess at different levels within the cognitive domain (i.e. different levels of Miller’s pyramid). The MCQ test predominantly assesses knowledge, whereas the simulation assessments assess performance (in this case application of knowledge in clinical situations, rather than technical competence in performance of a skill). Essentially, then, the simulation scenarios assess the same construct as the MCQs, but at a higher level. While not included in the simulation assessments for this research, it would also be possible, if desired, to assess non-knowledge-based constructs through the simulation assessments, for example communication and professionalism.

The correlation between the two methods of assessment suggests that simulation assessment could be considered as an alternative to the standard approach of MCQ tests, in terms of being able to assess similar constructs. Of course, any consideration of changing assessment method would need to take into account multiple other factors,
including all the psychometric properties (reliability and validity) evaluated in this study, as well as feasibility and acceptability.(11)

This part of the discussion has addressed Research Question 2- “What is the relationship between performance on simulation-based and MCQ assessments of cardiac emergency management?” The next chapter will go on to present the findings on Research Question 1E- “What is the content validity of each method of assessment?”
CHAPTER 18: CONTENT VALIDITY

This chapter will address Research Question 1E- “What is the content validity of each method of assessment?” and contribute to addressing Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?”

A. Multiple Choice Question Test

Qualitative analysis of the MCQ test has demonstrated that this assessment tool has adequate content validity, as both topics (AMI and APO) are addressed equally, all learning outcome categories are assessed, and there are no items that lack relevance to the topics. Determining the degree of content validity is more difficult and more subjective.

Learning outcomes were listed under categories and subcategories for the purpose of developing assessment tools with due attention to content. In an ideal assessment, all learning objective categories would be assessed by the same number of items, or the number of items per learning outcome category would be determined by the relative weighting of the learning outcomes. However, in clinical practice there is significant overlap between all aspects of patient care, and those areas listed as categories and subcategories for learning outcomes are perhaps more aptly termed clinical skill sets.

Figure 18.1 is a diagrammatic conceptualization of the relationship between the clinical skill sets. The two major categories of skill sets in this conceptualisation are patient assessment and patient management. There is an area of overlap between these, which includes prioritisation and re-evaluation. It should be noted that in in this conceptualization, the term re-evaluation replaces the subcategory patient progress,
which was listed in the learning outcomes. *Patient progress* is a term that is usually used in reference to patient outcome(s) in the context of a mid-to-long-term patient journey. On reflection, it was considered that the term *re-evaluation* better reflected an assessable *clinical skill set*, along with the other skill sets (history, examination, investigation, diagnosis, prioritization, treatment).

![Figure 18.1: Conceptualisation of the clinical skill sets](image)

This conceptualization possibly oversimplifies the complexity of the relationships between the skill sets. In practice, there is significant and frequent overlap, interaction and feedback between these areas. For example, findings on history will guide examination, and synthesis of the findings of history and examination guide investigation. The combination of history, examination and investigation findings inform a list of differential diagnoses to be considered. From this list, which must always include the most common and most critical diagnoses, the likelihood of each differential diagnosis is weighed by incorporating all available information. The process of narrowing this differential diagnosis list may involve obtaining further history,
performing further examination or requesting additional (perhaps more advanced or sophisticated) investigations.

Prioritisation and re-evaluation, while placed in the overlapping area between patient assessment and patient management in Figure 18.1, are more actually overarching considerations that are re-considered continuously throughout the patient journey. Re-evaluation is performed both consciously and subconsciously, and with a frequency that depends on the stability and severity of the clinical situation. In a resuscitation situation, the patient’s airway, breathing and circulation are very frequently re-assessed and, in some very critical situations, individuals may be allocated specifically to assess and manage each of these. Prioritisation is also ongoing, and initial evaluation and re-evaluation prompt re-consideration of prioritisation.

The very considerable gap between the highly interrelated, overlapping clinical skill sets and the categorisation of learning outcomes means that assessment of the distribution of items between the categories is not an accurate method to evaluate the degree of content validity. However, in the absence of any superior method, these findings must be considered.

The distribution of the MCQ items was skewed in favour of investigation and management categories, with eleven of the thirty items falling under each of these categories; there was approximately equal distribution of investigation and management questions for AMI and APO topics. In comparison, each of the other categories (history, examination, diagnosis, prioritisation and patient progress) was assessed by one to three MCQ items. Prioritisation was assessed for both of AMI and APO, but history, examination, diagnosis and patient progress categories were only assessed for either
AMI or APO.

The skewed distribution, in preference of items assessing investigation and management, is appropriate to the assessment form. It has been established that practical skills are best assessed by practical assessment methods, while knowledge is best assessed by written assessment methods, including MCQ tests. (11, 22) This was recognised in the development phase of the MCQ test, when it was decided that the same specific learning objectives used in construction of the simulation assessment tools were not able to be adequately assessed in the form of an MCQ test.

Examination skills are not well assessed by written forms of assessment, as they are essentially a practical skill; while MCQ items may be able to test knowledge of what examination should be performed in a given situation, or interpretation of examination findings, they are unable to assess examination skills directly. Similarly, while MCQ test items can assess ability to interpret history, they cannot assess the ability of a candidate to take a history, as this is a practical skill, rather than theoretical/knowledge-based competency. (11, 22)

Diagnosis, prioritisation and patient progress (re-evaluation) are skill sets that require the interpretation and synthesis of information from multiple sources, in a specific context; this requires both knowledge and problem-solving abilities. While attempts can be made to assess these skill sets using MCQ items, these items only test ability to interpret very concise and purposefully selected information, and select the (most) correct response from the alternatives provided. In this way, MCQ questions predominantly assess knowledge, rather than problem-solving ability. (102)
The reason that investigation and management were the focus of the MCQ test is because these are skill sets that can be assessed in this format. MCQ tests can be used to assess a candidate’s ability to select appropriate investigation(s) in a given situation. This parallels the situation in which a junior doctor has obtained history and performed an examination, and must then select the appropriate investigations. MCQ items may present this as a question of which test is most appropriate (or most accurate, most useful, etc) in the context of a clinical vignette; in this case the candidate must first interpret the situation to determine the purpose of the investigation. Alternatively, the item may be presented as a simple knowledge recall question of which investigation should be employed for a specified purpose.

In terms of assessing management, in this era of evidence-based, guideline-based practice, there is generally a best-evidence management choice for all common situations. An MCQ item is, therefore, able to assess knowledge of the appropriate medication to treat a given diagnosis, and may do this with some contextualisation.

Based on the qualitative assessment performed, the MCQ test developed for this research project has adequate content validity. However, the distribution of the content covered in this assessment tool favours assessment of the clinical skill sets of investigation and management. As discussed, however, the uneven distribution of items between the different learning outcome categories, in this case, is a reflection on the limitations of MCQ tests in medical education assessment; this is an important reason for considering alternative assessment methods.
B. Simulation Assessments

The learning objectives and simulation assessment sheets developed for this study were designed based on a traditional systematic and sequential approach and listed under corresponding headings: history, examination, investigation, diagnosis, management. (103) This approach is commonly taught in medical school, (103) and was determined to be the most appropriate given the clinical scenarios and participants involved. The simulation scenarios portrayed relatively stable patients who did not require immediate resuscitation, and for whom it would be appropriate to take a fairly comprehensive history and perform a focused examination prior to requesting investigations and initiating management. However, no marks were allocated on the simulation tool to the sequencing of the assessment and management tasks (with the exception of requiring re-assessment of the patient at some stage during the simulation).

The data demonstrate that both simulation scenario assessments covered each learning outcome (Tables 12.3 and 12.5). The spread of items across the learning outcomes was also fairly even, though there were a greater number of items assessing history, examination, investigation and management. While there were fewer items that assessed diagnosis, patient progress and prioritisation, compared to the other learning outcome categories, this was clinically appropriate. The reason that this is considered appropriate is that, diagnosis, prioritisation, and assessment of patient progress would realistically only be addressed once in a ten minute time period in clinical practice, unless the patient is very unstable. The even spread of items across the learning outcome categories reflects the fact that simulation, as a practical assessment tool, is able to assess all aspects of practical skills (AMI and APO emergency management). (11, 22)
The data demonstrate that each specific learning outcome was covered by at least one assessment item, indicating good content validity (Tables 12.4 and 12.6). The only possible issue identified in these tables is that some assessment items assessed more than one learning outcome and, in some cases, the only item that assessed a given learning outcome was not specific to that outcome; i.e. covered more than one outcome. While it is true that execution of a particular management step or checklist item may reflect attainment of more than one learning outcome, it would be preferable for each item to address a single learning outcome, so that assessors are better able to gauge attainment of each outcome (this is discussed further in Chapter 20: Formative Validity). However, since simulations by definition attempt to recreate the complexity of clinical practice, some complexity in overlap of learning objectives is to be expected.

Based on the qualitative assessment performed, the simulation assessments developed for this research project have very good content validity and relatively even coverage of the learning outcomes/learning outcome categories. This is a reflection of the fact that simulation assessment is practical assessment method, suitable for assessment of the practical skills of cardiac emergency management. (22) Essentially, since the simulation scenarios were able to closely mirror the realities of clinical situations, the assessment tools were able to achieve very good content validity.

C. Comparison of the Multiple Choice Question Test and Simulation Assessments

Qualitative evaluation of the content validity for the MCQ test demonstrated that the test had adequate content validity, but uneven distribution of the content covered in favour of the clinical skill sets of investigation and management. As discussed, this bias towards addressing areas that are more knowledge-based is a limitation of written
assessment methods, especially multiple choice question tests. (11) The simulation assessments had very good content validity and were able to more evenly address the different learning outcome categories. Again, this was attributable to the fact that simulation is a practical assessment method designed to replicate clinical practice, and therefore is able to assess all facets of clinical management topics. (22)

In comparing content validity and content coverage of the two assessment methods, it is also pertinent to consider the number of items used to assess the content, and the time allowed for each assessment, sometimes termed testing efficiency. (59) The MCQ test used 30 items, with a maximum attainable mark of 30, to assess both AMI and APO management. Testing time was 30 minutes (maximum time allowed). Some additional time was required for set-up (approximately 2 minutes per sitting), and manual marking of each participant’s answer sheet (approximately 2 minutes per participant).

- Each simulation assessment test used 26-27 checklist items and a global assessment score (GAS), with a maximum attainable mark of 30, to assess either AMI or APO management. Testing time was 10 minutes, plus 5 minutes debrief. Some additional time was required for set-up/re-setting between simulations (approximately 5 minutes total per participant), because the two simulations were run alternately.

On an individual level, the MCQ test assessed the content of both simulation assessments combined (AMI and APO management), in the same time that would be taken to run both simulation assessments, i.e. 30 minutes was required to assess both topics, using either assessment approach. However, the number of items used to assess the topics was different for the different assessment methods. The MCQ test assessed
both topics with 30 items, whereas the simulation assessment assessed each topic with just under 30 items. As such, approximately the same amount of numerical information is obtained about the examinee in fifteen minutes of simulation assessment time, as is obtained in thirty minutes of MCQ assessment time. This implies that in this case, the simulation scenarios had better testing efficiency, on an individual level, than the MCQ test. (59)

The other pertinent difference, which is important in considering assessment feasibility, is the time taken for assessment of a cohort. (11) While the MCQ test can be delivered to multiple examinees at the same time, the simulation assessment, in its current form, is delivered for individual candidates. Therefore, if there was a large number of candidates being assessed using the simulation assessment tool, the time (and resources) required would be very much more than for an MCQ test. For example, if a cohort of one hundred examinees were to be assessed using an MCQ test, the estimated total examination time for the entire group would be 240 minutes (30 minutes for the MCQ test, plus ten minutes set up for a larger cohort, plus two minutes per participant for marking). Whereas using a simulation assessment approach, with one simulation assessment per participant, the time for the entire group would be 2000 minutes (15 minutes per participant for the simulation including debrief, plus five minutes per participant for set-up/re-setting between simulations).

These examples are based on the limited resources we had available for this research project. If two simulations could be run concurrently, this time would halve to 1000 minutes, and time taken would decrease further if multiple simulations could be run concurrently. However, if more than one simulation scenario was required per
examinee, the time would increase linearly with the number of scenarios. This would also apply if there were a larger number of MCQ items. Overall, though, it is clear that use of simulation-based assessments to assess an entire cohort requires significantly more time and human resources compared to MCQ-based assessment.

This chapter has addressed Research Question 1E- “What is the content validity of each method of assessment?” and contributed to addressing Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?” The next chapter will discuss Research Question 1A- “What is the construct validity?”
CHAPTER 19: CONSTRUCT VALIDITY

This chapter will provide an evaluation of Research Question 1A- “What is the construct validity?” and contribute to evaluation of Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?”.

A. Multiple Choice Question Test

i. Internal Consistency

For good internal consistency, Cronbach’s alpha should be above 0.70.(50, 52, 88)
Considering that Cronbach’s alpha increases with a larger number of test items, an upper limit of 0.90 has been recommended by some researchers, as values higher than this suggest that there are redundant items which do not add value.(52, 88) However, other authors suggest that higher values (≥ 0.90) are desirable, and that some redundancy in items is necessary.(50, 52)

Cronbach’s alpha for standardised items for the entire MCQ test is 0.64 (Table 13.1). This value, while close to the commonly accepted minimum value for internal consistency (0.70), is lower than is desirable, and indicates that the MCQ test has insufficient internal consistency. The implication of this finding is that the items on the MCQ test are insufficiently correlated with each other, and may assess different constructs.(50)

As previously discussed, the number of items on the test significantly affects Cronbach’s alpha, and it is likely that the relatively small number of items on the test (30 items) also contributed to the low coefficient alpha value. This value would be expected to improve with the addition of further items, so long as the additional items correlate well with the existing items.(50, 52, 88, 89) However, it is preferable to
maintain the same number of items and improve the coefficient alpha by excluding poorly correlated items, and replacing them with highly correlated items. (89)

The purpose of showing Cronbach’s alpha for each MCQ item when all items are considered as a single assessment tool was to identify those items with low correlation, which decrease the internal consistency of the instrument (Table 13.2). In interpreting the results shown, both columns are of interest.

Firstly, items with a corrected item-total correlation that is close to zero, or negative, correlate poorly with the other items (the majority of the items). (88) However the column that shows Cronbach’s alpha if the item is deleted, gives the best indication of which items are contributing negatively to the overall internal consistency. The items for which this value is higher than 0.64 should be deleted, and replaced with items that are more highly correlated. (88, 89) Specifically, items 2, 12, 26, 27 and 30 are identified as items that, if removed, would improve the overall test coefficient alpha. However, it should be noted that for all of these items, Cronbach’s alpha would only very marginally improve upon their removal. Therefore, it would be more practical to include additional items with high item-total correlation. (89)

Internal consistency was also calculated for the two different components of the MCQ test, that is, AMI items and APO items. Cronbach’s alpha for standardised items for the AMI component of the MCQ test is 0.68. This approaches the recommended minimum value for internal consistency, and is slightly higher than the coefficient alpha for the entire test (inclusive of both AMI and APO components). However, it still indicates inadequate internal consistency. Again, the small number of items negatively impacts the coefficient alpha, and this effect is compounded when examining the smaller
components within an assessment tool. Interpretation of Table 13.3 shows that items 2 and 15 have low item-total correlation, and actually, removal of either (or both) of these items would raise the coefficient alpha for the AMI component to an acceptable value of 0.70. As previously discussed, an alternative would be to include additional items.

Cronbach’s alpha for standardised items, for the APO component of the MCQ test, is extremely low, implying that this component of the MCQ test completely lacks internal consistency. As such, Table 13.3 is difficult to interpret, since there is overall poor correlation between items.

Given that items were purposely written to test the same construct (management of APO) and adequate content validity has been proven, it is important to consider other possible reasons for such a low coefficient alpha value for the APO component of the MCQ test, apart from items assessing constructs other than APO management. As previously mentioned, a low coefficient alpha indicates heterogeneity, and is also decreased if there are fewer items. (50, 88) While adding correlated items would improve the coefficient alpha value, this is not a viable (or educationally sound) solution when the internal consistency is so low to begin with, as it does not address the underlying problem.

A possible explanation for the very low coefficient alpha is that there is, in fact, significant heterogeneity in the concepts that make up APO management. Whereas AMI is a distinct, defined pathological process, APO can be a manifestation of a range of different pathologies, including AMI itself. In fact, any process that impairs the ability of the left heart to effectively pump blood into the systemic circulation, as well as any process that limits the ability of the body to eliminate excess fluid, can result in APO.
As such, the topic of APO is very broad, and in order to assess APO management, items covering a broad range of subtopics must be included in the MCQ test. Therefore, it is possible that, in assessing APO, multiple smaller constructs are actually being assessed. If participants performed differently on each of these smaller constructs, then this would explain the significant heterogeneity in examinees’ performance on the APO items.

Ideally, a pilot run of the assessment would have been undertaken, and, recognising the low coefficient alpha, additional items would have been added to the assessment, and the coefficient alpha re-assessed. This process would then continue until the coefficient alpha is adequate; in some cases multiple rounds of piloting are required.(52, 70, 79, 88) Unfortunately, a lack of access to suitable volunteers (apart from potential research participants who, if involved in piloting, would then not be able to participate) prevented full-scale piloting.(79)

It is important to remember that internal consistency is specific to a single administration of the assessment and will differ with different cohorts/sittings of an assessment. As such, the internal consistency calculated in this research project relates only to this single execution of the MCQ test, and should be calculated each time the assessment is administered.(52)

**ii. Contrasted-Groups Approach**

MD2 and HMO2 groups were chosen to represent ‘novice’ and ‘expert’ groups respectively, since these were the two groups of participants with the most difference in regards to experience and seniority. MD2 medical students at Northern Clinical School are in their first year of clinical (hospital) training and, while they have had classroom-based teaching on a range of cardiology topics, including common pathologies, during
their first (non-clinical) year, they have had very little, if any, clinical exposure to cardiology patients/presentations. Conversely, HMO2 doctors have had a full year of experience working in the clinical setting, in addition to completing their medical school (pre-clinical and clinical) training. All interns in Australia are required to complete at least one of each of a medical, surgical and emergency rotation, in order to qualify to upgrade from probationary to general medical registration. As such, all HMO2s would be expected to have some experience in managing patients with cardiac conditions and patients with emergency presentations for chest pain and dyspnoea.

Consideration was given to re-grouping classifications into broader comparison groups. In particular, it was considered that it might be useful to compare medical students to junior doctors. During initial phases of statistical analysis, comparisons were made between the results of medical students and junior doctors, which demonstrated no difference in performance between these two groups. This may be attributable to the lack of experience of the intern group, in that the research project was undertaken in April/May, by which time the interns had only been working for approximately 3 to 4 months, with close supervision, and may not yet have had any experience working in medical or emergency rotations. Perhaps there would have been observable differences between the medical student and doctor groups later in the year, once the interns had gained more experience working as junior doctors beyond medical school.

While the comparison between medical students and junior doctors would be of interest, it was determined that combining levels of experience would not be an appropriate method for determining construct validity. As previously mentioned, the contrasted-groups approach to determining construct validity depends on there being a large difference in expected performance of the two groups on the construct being
measured. (75) If medical students were compared to junior doctors, then there would be only one year difference in classification/experience between some participants in the two comparison groups, so the expected difference in performance would be small. However, a much larger difference in performance between second year medical students (MD2) and HMO2 level junior doctors would be expected, with a difference in experience of four years.

Comparison of the novice and expert groups demonstrated a statistically significant difference in MCQ scores between the groups. This indicates that the MCQ test was able to differentiate between novices and experts in the construct being tested, that is, management of cardiac emergencies. As such, the construct validity of this assessment is confirmed. (73, 75, 76)

Despite statistical significance in performance between novice and expert groups being determined by the Mann-Whitney U test, it is interesting to note that the difference in mean scores between MD2 and HMO2 participants is approximately equal to one standard deviation. This means that some MD2 students performed as well as, or better than, some HMO2 doctors (Figure 13.1). While this does not negate the construct validity of the test, it is important to interpret the implications of this observation.

There are many explanations for the observation that some of the MD2 students outperformed some of the HMO2 doctors. For example, it is possible that there was some self-selection bias; the most motivated and best MD2 students may have chosen to participate in the research project, perceiving the project as an opportunity to get a head start in clinical practice, whereas it may have been predominantly the weaker HMO2 doctors who felt a need to gain extra experience and feedback through simulation.
(However, it is unlikely that *only* the weaker HMO2 doctors participated, since some of the best simulation assessment scores were those of HMO2 participants). Another explanation is that the MD2 participants may have put in more effort than the HMO2 participants; despite the knowledge that performance on this research project would not contribute in any way to university assessment, MD2 students may have been keen to gauge their level of performance, as an indicator of their progress. HMO2 participants, having no formal assessments to prepare for, and having passed internship, would be less likely to attribute any importance to their performance on this assessment. Another possibility is that MD2 students had better *exam technique*, given that they were still regularly sitting formal exams, whereas most HMO2 participants would not have undertaken a formal exam in the approximately 16 months since they completed medical school.

The overlap in performance between ‘novice’ and ‘expert’ groups could also call into question the definition of these two groups. Ideally, novice and expert groups would be more widely separated than the four years of experience that separates HMO2 participants from MD2 participants. MD2 students are true novices to clinical practice but, while comparatively more experienced, HMO2 doctors are by no means true experts. If only construct validity was being assessed, a true expert group would be used, such as emergency physicians or cardiologists. However, this was outside the scope of this research project and would not necessarily be appropriate for examining the other psychometric properties of the assessment tools.

It is important to note that previous studies have also demonstrated that, on checklist-based assessments, there may not be significant differences in performance between participants of different levels. (59) This may suggest that the reason for the lack of
difference is accounted for by inherent characteristics of a checklist-based simulation assessment, rather than necessarily reflecting the characteristics or performance of the groups. Regardless of the reasons why some of the MD2 participants outperformed some of the HMO2 participants, it is important to emphasise that, overall, the HMO2 participants performed significantly better than the MD2 participants on the MCQ test, as expected.

iii. Comparison Across All Groups

The Kruskal-Wallis comparison of performance across all classifications produced somewhat unexpected results. Progressive improvement in performance from most junior to most senior participants was not demonstrated (Figure 13.2). The mean score was higher in MD3 students than in MD2 students, and higher in HMO2 than intern doctors, but performance of MD4 and intern groups was lower than that of MD2 students. Kruskal-Wallis comparison demonstrated that there was a significant difference across all groups, but the Dunn-Bonferroni post-hoc test clarified that the only significant difference between any two specific groups was between intern and HMO2 groups. The difference in mean MCQ score for these two groups was relatively large.

The contrasted-groups approach was prospectively identified as the appropriate test to determine construct validity. (73, 75, 76) As such, the finding of inconsistent differences across all classifications does not negate the fact that the construct validity of the MCQ test was proven by comparison of novice and expert groups. (75) It may, however, suggest that the MCQ test is unable to accurately distinguish between performance of examinees at a similar level. (75) An alternative explanation is that the part of the
construct and content that is examinable in the form of an MCQ test, does not vary significantly across the classifications MD2 to intern.

As previously discussed in *Chapter 18: Content Validity*, the MCQ test predominantly assessed investigation and management skills, and this is appropriate to the form of the assessment. Similarly, given the form of the assessment, the specific *construct* being tested by the MCQ test may actually be more specific than cardiac emergency management and, perhaps could be better defined as *theoretical knowledge of cardiac emergency management*. The findings of the Kruskal-Wallis comparison across all classifications are consistent with, and could be interpreted as, a lack of difference in these more specifically defined aspects of content and construct.

This more specific interpretation of results may explain why there *is* a significant difference in performance between intern and HMO2 groups. Medical students learn about cardiac emergencies and the theoretical underpinnings of their management. Then the intern year provides the first opportunity to actually practice medicine, that is, make clinical judgments and decisions. It seems logical, then, that knowledge of how to manage cardiac emergencies would improve with the opportunity to practise these skills over the intern year. It is also important to note that, since data collection was undertaken in April/May, interns had only been working for three to four months, such that their clinical experience would be more similar to a final year medical student. Results may have been different if the simulations had been performed at the end of the year, after the intern cohort had gained more clinical experience.

It is also necessary to consider that the results obtained may have been impacted by sample sizes. The sample sizes of some classification groups were very small: three
classifications comprised of less than ten participants, the smallest sample size was for the MD4 classification, with only three participants. The smaller number of participants decreases the accuracy of the results, and the ability to draw conclusions that relate to the larger cohort from which these participants were drawn.(50)

Further compounding this issue is the possibility of selection bias. While there was no bias in the distribution or content of recruitment material, it is possible that there was some self-selection bias, as previously discussed. That is, participants who chose to participate were those who, based on the recruitment material, felt either more confident with the content (cardiology or emergency topics) or assessment method (simulation, which was advertised as the focus of the project). It is also possible that there may have been differential selection bias between students and junior doctors. For example, the opportunity to participate in a cardiac emergency simulation may have attracted the more confident students and the less confident junior doctors, for example, those junior doctors who feel they have had inadequate exposure to cardiac emergency management.

Another important finding was that the comparison between MD2 and HMO2 groups yielded different results with the two different statistical tests used: the Mann-Whitney U test and the Kruksal-Wallis comparison with Dunn-Bonferroni post-hoc. The former of these tests found a statistically significant difference between MD2 and HMO2 scores, but the latter found the difference not to be significant. This may be explained by the fact that the Dunn-Bonferroni post-hoc method, when applied to the Kruskal-Wallis test, corrects for multiplicity/multiple comparisons. This correction takes into account the number of comparisons being made and is correctly applied when searching for significance where multiple comparisons, variables or hypotheses are considered and therefore the likelihood of finding a significant result is falsely increased. This is
corrected for by mathematically determining a corrected significance level. (83, 105, 106)

The Dunn-Bonferroni post-hoc method has been correctly applied to the pairwise Kruskal-Wallis test, where the intention is to determine whether there is a significant difference between the performance of any two groups. Conversely, a Dunn-Bonferroni correction was not required when specifically comparing the performance of the predetermined ‘novice’ and ‘expert’ groups on a single dependent variable.

B. Simulation Assessments

i. Internal Consistency

Cronbach’s alpha for standardised items for Simulation Scenario 1 was unacceptably low, meaning that there is insufficient correlation between the items. However, it is important to note that the literal interpretation is that there is low correlation between participant performance on the items. Correlation is analysed based on how participants perform on the assessment, rather than the actual content of the items. As such, it is possible that this result could reflect heterogeneity in the level of performance of participants across different aspects of a single construct. This interpretation becomes less likely when there are more participants, since some participants would be expected to perform consistently across a single construct.

We must then consider the effect of each individual checklist item on the assessment tool on the measure of internal consistency. For Simulation Scenario 1, there is only one item (Item A3a), which, if removed, would improve the overall test coefficient alpha value. This suggests that apart from this one item, there are no specific items (or
groups of items) that are negatively impacting the internal consistency, but rather there is a general lack of correlation between performance across all items. It is also important to note that if additional items were added to the assessment checklist, this would be expected to improve the coefficient alpha value, so long as the new items were correlated to the existing items.

Cronbach’s alpha for standardised items for Simulation Scenario 2 (checklist component), which assessed performance on an APO scenario, had complete lack of internal consistency; there is no demonstrable relationship between performance across different items on the assessment. As for the APO component of the MCQ test, this may be partly attributable to the significant heterogeneity in the concepts that make up APO management. Specifically, in relation to Simulation Scenario 2, there was a level of complexity that required understanding of a broad range of content/constructs, even if this was not the focus of the assessment.

In Scenario 2, the acute deterioration in the patient’s condition was due to APO, however APO was precipitated by atrial fibrillation with rapid ventricular rate, and excessive intravenous fluid was also a contributing factor. Furthermore, the worsening atrial fibrillation itself was secondary to infection. Another complicating factor was that the patient’s primary complaint was dyspnoea, which is a common undifferentiated symptom across a range of body systems (though predominantly cardiovascular and respiratory), with a large range of differential diagnoses. Therefore, to successfully assess and manage the patient in this scenario requires knowledge and skills across a range of topics/constructs.
Therefore, the most likely explanation for the extremely low coefficient alpha for Simulation Scenario 2 is that there are confounding factors influencing performance on the assessment. While the low correlation suggests that the test items are not related, we know this is not the case, as the assessment is based on a single clinical scenario, and the items correspond to a checklist of expected actions in management of APO. The confounding factors that are influencing the statistical measure of internal consistency are: firstly, aptitude across other topics/constructs that are involved with the scenario but which are not the focus; secondly, the undifferentiated nature of the primary problem, and the breadth of possible diagnoses.

Given the very low overall coefficient alpha value for Simulation Scenario 2, it is difficult to properly interpret the effect of individual checklist items on the overall internal consistency measure. However, there are two findings to highlight; the first of these is that there are a few items with relatively high corrected item-total correlation (this is relative to the overall correlation)- specifically items A2b, A4a, A4c and A4d (Table 13.9). It is important to note that most of these items fall within a group, in that they were all items assessing investigation of APO. High correlation between these items and the overall score may suggest that these items were part of a learnt list of investigations for AF/APO, and that those candidates who were proficient and confident at managing this clinical situation had mastered and were able to recall this list. However, it is difficult to interpret the significance of higher item-total correlation for specific items in the context of the overall very low coefficient alpha value for Scenario 2.

The second finding was that there were some items with significantly negative (not close to zero) corrected item-total correlations, suggesting that there was an inverse
relationship between performance on these items and overall performance on the checklist component of Simulation Scenario 2. That is, if a participant performed well on this item, they were less likely to perform well overall and vice versa. A possible explanation for this is that participants of lower overall ability might be more likely to focus on very specific aspects of assessment and perhaps a single pillar of management, rather than completing a more broad assessment and employing a comprehensive management strategy.

As discussed in relation to the MCQ test, the finding of low coefficient alpha values on piloting would usually prompt changes to be made, and multiple rounds of piloting may be required to achieve an adequate coefficient alpha.(52, 70, 79, 88) However, as previously discussed, full-scale piloting of the assessment tools was not possible for this research project, without compromising recruitment.(79)

ii. Contrasted-Groups Approach

The analysis for construct validity using the contrasted-groups approach was limited by the number of participants. Specifically, only one MD2 participant was allocated to Simulation Scenario 1, so the results for MD2 participants on this assessment are not a reflection of a group, but rather an individual. As such, it was decided that the MD3 group should be used instead of the MD2 group for this particular comparison, as this would be the closest to a group representation of ‘novices’. There was no statistically significant difference in performance between the ‘novice’ and ‘expert’ groups for Simulation Scenario 1 (Table 13.4 and Figure 13.3). As such, by the definition used for this statistical approach, Simulation Scenario 1 lacks construct validity. It is pertinent to consider that the use of a slightly more experienced ‘novice’ group (MD3) may have influenced this outcome.
Comparison of the novice and expert groups demonstrated a statistically significant difference between the groups for overall score on Simulation Scenario 2. This indicates that Simulation Scenario 2 was able to differentiate between novices and experts in the construct being tested, that is, management of APO. The difference in performance of the two groups was also significant for the two components of the simulation assessment (checklist and GAS, Table 13.4). As such, the construct validity of Simulation Scenario 2 is confirmed, based on novice/expert differentiation criteria.(73, 75, 76)

**iii. Comparison Across All Groups**

The descriptive statistics and Kruskal-Wallis comparison of performance across all classifications for Simulation Scenario 1 gave somewhat unexpected, though not particularly significant, results. Essentially, there was no identifiable pattern in the performance of participants across the range of classifications, and all mean results were within a very small range (1-2 standard deviations). Importantly, the Kruskal-Wallis comparisons for the total score, and component scores, for Simulation Scenario 1, demonstrated no significant difference across all groups.

These findings suggest that the Simulation Scenario 1 assessment tool was unable to accurately distinguish between performance of examinees at a similar level.(75) As previously discussed in *Chapter 17: Participant Characteristics and Assessment Results*, participant results for Scenario 1 were generally high, suggesting that this assessment may have been of insufficient difficulty; if this was the case then this would explain the inability of the test to discriminate between participants of different levels.

The results for Simulation Scenario 2 are more in line with expectations. Mean total
scores progressively increased with level of experience across all groups (except MD4, for which no participants were allocated to Scenario 2). This pattern was similar for checklist and GAS components, with there being some exceptions to the pattern: interns performed slightly better than the HMO2 group on the checklist score, and MD3s performed slightly better than interns on the GAS. The Kruskal-Wallis comparisons for total score and component scores for Scenario 2 were all significant. However, pairwise comparison across all classifications, using the Dunn-Bonferroni correction, demonstrated significant differences (adjusted significance value < 0.05) between the MD2 and HMO2 groups only, for both component scores and overall score.

These findings suggest that the Simulation Scenario 2 assessment tool was reliably able to distinguish between the most junior and most senior examinees, in keeping with the findings of the Mann-Whitney U test comparing ‘novice’ and ‘expert’ participants. While there was not statistically significant differentiation between participants of more similar experience levels, a finding that is also consistent with previous studies,(59) the pattern of scores was consistent with a test that is able to make this differentiation. Therefore, it is pertinent to consider reasons that statistical significance was not reached. The most likely limiting factor was the small number of participants of each classification (and lack of any MD4 participants allocated to Scenario 2). With small numbers, individual performance has a greater effect on mean scores and therefore the sampling error becomes more prominent.(50) As such, it is more difficult to prove that any findings are due to an inherent difference in the groups, rather than chance; i.e. it is more difficult to reach statistical significance.(50) Nonetheless, the findings for Simulation Scenario 2 do give some promising evidence of ability to differentiate across a range of classifications.
C. Summary of Construct Validity

The important findings in regards to construct validity were:

1. The MCQ test, as a whole, had low (inadequate) internal consistency.

2. The AMI component of the MCQ test had low (but almost adequate) internal consistency, which could be improved to an acceptable level with removal of one to two items.

3. The APO component of the MCQ test had very low internal (inadequate) consistency, and this may be explained by the broad nature of APO as a process.

4. The MCQ test had statistically significant construct validity, as evaluated using the contrasted-groups approach.

5. Performance on the MCQ test does not consistently improve with the experience level or seniority of the participants across all classification.

6. The assessment tool for Simulation Scenario 1 had low (inadequate) internal consistency.

7. The assessment tool for Simulation Scenario 2 had no evidence of any internal consistency.

8. Simulation Scenario 1 had low construct validity, as evaluated using the contrasted-groups approach.

9. Simulation Scenario 2 had statistically significant construct validity, as evaluated using the contrasted-groups approach.

10. Performance on Simulation Scenario 1 (total score) did not consistently improve with the experience level/seniority of the participants across all classification.

11. Performance on Simulation Scenario 2 (total score) consistently improved with the experience level/seniority of participants, however, there was only a statistically significant difference in performance between the most junior (MD2) and most senior (HMO2) groups.
D. Comparison of the Multiple Choice Question Test and Simulation Assessments

Overall, the MCQ test had higher internal consistency than the simulation assessments. The reason for the lower internal consistency of the simulation assessments, as discussed previously, could relate to the heterogeneity and complexity in the components that make up clinical management of cardiac emergencies. There is less confounding for MCQs since each item, at least, is assessing a single construct. It is also clearer from a written stem, compared to a clinical scenario, what the construct being assessed is: for example, a vignette must give sufficient information to be able to logically differentiate the pathology in question; whereas a clinical scenario is potentially more ambiguous.

Construct validity was also evaluated using the contrasted-groups approach. This approach proved the construct validity of the MCQ test and Scenario 2, but not Scenario 1. Comparison across all classifications failed to show a consistent improvement in performance on the MCQ test or Scenario 1 with increasing experience. There was consistent improvement across classifications for Scenario 2; however, this was not statistically significant (except for the difference between MD2 and HMO2 participants).

Overall, based on the combination of measures, the MCQ test had the highest construct validity of the three assessment tools, though the level of construct validity was still poor. Given the different findings for the two simulation scenarios, it would appear that the construct validity was less a property of the assessment type/format, but rather related more to the individual assessment tool.
This chapter has provided an evaluation of Research Question 1A- “What is the construct validity?” and has contributed to evaluation of Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?”. The following chapter will address Research Question 1B- “What is the formative validity?”
CHAPTER 20: FORMATIVE VALIDITY

This chapter will present the research findings on Research Question 1B- “What is the formative validity?” and contribute to answering Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?”

A. Multiple Choice Question Test

i. Quantitative Component

The mean value for each MCQ item is a measure of the actual difficulty (as opposed to theoretical or expert-determined difficulty). (98) A lower mean value implies a higher level of difficulty. As such, Table 14.1 demonstrates that MCQ item 20 was the least difficult, with all participants answering correctly; Item 8 was the most difficult, with only approximately 45 per cent of participants answering correctly.

It is important to note that in an MCQ test with positive marking (i.e. marks allocated for each correct answer, no marks subtracted for an incorrect answer), the score achievable by chance depends on the number of options given. (98) The rationale for choosing to construct an MCQ test with three options per item has been discussed earlier. Since the MCQ items each had three options, a score of 33% (or total score of 10/30) would be achievable by chance. All items have mean values above 0.33, suggesting that all participants had some knowledge of the topics being tested and made some genuine effort to answer based on this knowledge.

Medium item difficulty for an MCQ test with three options per item is approximately 0.67. (98) In order to best discriminate between examinees and provide a high level of reliability, there should be some items across each level of difficulty, with a concentration of questions of slightly higher than medium difficulty. (98) The data
demonstrate that there was a spread of difficulty in the items (Table 14.2 and Figure 14.1). While there was only one item with difficulty in the range 0.41-0.50, and relatively few items in the range 0.51-0.60, the overall spread of results (discussed previously), suggests that there was sufficient discrimination between the higher-achieving examinees. Only one candidate scored 30/30 on the test, and the next highest score achieved was 28/30, also only attained by one candidate.

The proportion of items with low difficulty was large, suggesting that the design of this assessment tool better discriminates between lower-performing examinees, and/or that the test is of insufficient difficulty. (98) This is supported by the distribution of results, discussed in Chapter 17. The lowest score attained on the MCQ test was 16/30, therefore all students answered at least half the items correctly.

This particular assessment was designed to be undertaken by examinees of a very wide range of levels (MD2 to HMO2), so it was considered necessary to include questions with a wide range of difficulties, and to include a sufficient number of questions of low difficulty, so as to be able to discriminate between all examinees. It is important to note that while it has been determined that discriminating ability of an assessment improves with items of slightly higher than medium difficulty, full evaluation of individual item discrimination requires additional calculations (e.g. quintile score groups, correct response curve, point biserial correlation), which are outside the scope of this research project. (98)

As previously discussed, learning outcome categories were unevenly represented in the MCQ test items; this must be borne in mind when considering the results presented in Table 14.3. The most important finding from this table is that for all learning outcome
categories that were assessed by more than one item, the items were of different levels of difficulty, thereby allowing differentiation of candidates’ performance in each of these learning outcome categories. The investigation category, in particular, was assessed by items of a range of difficulties, both for AMI and APO. APO management was assessed by items of low and high difficulty, but no items of medium difficulty.

Inclusion of items with a range of difficulties, in theory, improves the formative validity of an assessment tool, since the examiner/educator is better able to assess candidates’ level of performance for each component and therefore provide more useful feedback. (107) For example, if an examinee correctly answered items across a range of difficulties in the AMI investigation category, but was only able to correctly answer items of low difficulty in the AMI management category, then this may suggest that AMI management is an area for improvement for this individual.

In general, items of low difficulty are not particularly useful, as almost all examinees will answer these questions correctly, thereby decreasing the ability of the assessment tool to discriminate between examinees. Similarly, if all items for a given category are of high difficulty, then only the best performing candidates for this category can be determined, as all other candidates will answer these items incorrectly. (107) Ideally, most items on an MCQ assessment should be of medium difficulty. (98) Items of low difficulty should not be included, or only be included as ‘warm-up’ questions. (107) However, in the case of the MCQ assessment used for this research, some items of low difficulty were intentionally included due to the wide range of examinee experience (MD2 to HMO2), with the intention of providing some encouragement for the more junior examinees. Some items of high difficulty were included to discriminate between the highest achieving examinees. (107)
So far, the interpretation of the results presented has focussed on how performance on items of differing difficulty can provide information on individual students’ performance. However, this data can also be used to provide information on group performance and, therefore, guide further teaching, which is another purpose of formative assessment. While the mean performance on an individual item or an entire assessment tool is taken as a measure of actual difficulty, it can also provide an indication of class performance. Therefore, another interpretation of the same data is that, since there were a high proportion of items with mean score 0.91 to 1.0, there is a high proportion of the content which the cohort (entire cohort, not any particular classification) has mastered. In addition, it would appear that in general, candidates were familiar with investigation of APO. However, the data suggests that there are some aspects of APO management that are well understood, and other areas that are poorly understood. If this data was acquired in the context of formative assessment for an educational program, this information would be useful to guide further teaching. The data could also be further divided to determine if there were particular areas in which specific groups (classifications) were performing particularly well, or particularly poorly.

The results of the Mann-Whitney U test, shown in Table 14.4, indicate that there were two items on the MCQ test (items 8 and 9) that were reliably able to differentiate between ‘novice’ and ‘expert’ examinees. For both these items, there was a statistically significant difference in performance between the MD2 & HMO2 groups. However, it could be argued that multiple comparisons could confound the results, and therefore it would be appropriate to apply a Bonferroni correction in interpretation of the results. This concept has been explained previously in Chapter 19: Construct Validity. In this case, the Bonferroni correction is applied by modifying the
Anastasia Vlachadis Castles
21544591

threshold value at which a finding is considered statistically significant. In this case, 0.05 is used as the pre-determined value for statistical significance for single comparisons. Since comparisons are being made for n different items, the significance level, \( \alpha \), is determined by:

\[
\alpha = \frac{p}{n}
\]

Where:
- \( p \) is the predetermined significance level (in this case \( p = 0.05 \))
- \( n \) is the number of comparisons

Therefore:

\[
\alpha = \frac{0.05}{30}
\]

\[
\alpha = 0.0017
\]

Therefore, the adjusted significance level for multiple comparisons being made on this thirty-item test, is \( \alpha = 0.0017 \).

This means that if the results of the Mann-Whitney U test are adjusted for multiplicity, there is no statistically significant difference between ‘novice’ and ‘expert’ performance for item 8 or item 9 and, therefore, there are no items on the MCQ test that are able to reliably differentiate between these two groups.

The results of the Kruskal-Wallis test, presented in Table 14.5, indicate that item 25 was the only item on which there was a significant difference in performance between participants across the spectrum of classifications. However, Table 14.6 provides more specific information, with pairwise comparisons for each classification. It is important to note that the results of Table 14.6 already include a Dunn-Bonferroni correction for multiplicity. That is, instead of determining a corrected \( \alpha \) value at which there is considered to be statistical significance (as was calculated for the Mann-Whitney U
test), the p value itself is corrected for multiplicity and quoted as an *adjusted significance*.

Therefore, the important finding from Table 14.6 is that there is a statistically significant difference in performance between the following groups:

- HMO2 and MD4
- Intern and MD4
- MD4 and MD3
- MD4 and MD2

It would be expected that the statistically significant difference between performance of MD4 group and the two lower classifications (MD2 and MD3) would reflect the superiority of MD4 group. However, comparison of the mean item performance values for each classification, shown in Table 14.5, demonstrates that the mean item performance was the lowest for the MD4 classification and the same for all other classifications. As such, the only information provided from the Kruskal-Wallis test is that the MD4 group performed particularly poorly on item 25, and that no other item could differentiate between the different classifications.

As previously discussed, and discussed in more detail in *Chapter 25: Limitations*, small sample sizes make interpretation of results for subgroups of the entire study population difficult (e.g. analyses of the performance of different classifications). The effect of small sample sizes was taken into account in the selection of statistical analysis methods. For example, non-parametric tests were used to compare classification groups, noting that it is not possible to have a true normal distribution of results for comparison groups with as few as three participants (as in the MD4 group). However, choice of
appropriate statistical tests does not mitigate the fact that a small numbers of participants in each classification make it very likely that sampling errors will occur. That is, that the statistics obtained based on the results of the sample population for any cohort/classification will not represent the population from which the sample is drawn. (50)

In particular, there was a very low number of MD4 participants in this research project (n = 3), such that it is very unlikely that these three participants are an accurate representation of the entire MD4 cohort at Northern Clinical School. In addition, a smaller sample size (overall, and within groups) means that an individual’s performance has a greater impact on the overall results. This is especially true when examining the performance of classification groups with small numbers of participants. This may explain the unusual finding of the poorer performance of the MD4 group on MCQ item 25. If this result is analysed more carefully, we find that this poorer performance actually represents a single MD4 participant answering this item incorrectly.

ii. Qualitative Component

Figure 14.2 shows the results of the qualitative analysis of the MCQ feedback document (Appendix 4: Multiple Choice Question Test Feedback), listed under headings that describe the important aspects of feedback: form, type, quantity and quality. (91, 92)

Another important aspect of feedback is timing, that is, when examinees receive feedback in relation to the assessment. The timing of feedback was not considered in the qualitative analysis for formative validity because the timing of feedback in this case was dictated by the need to maintain the integrity of the research project. It was considered necessary for feedback on the MCQ test to be withheld until all participants
had completed the MCQ test, so that participants could not discuss the correct answers, which were provided in the MCQ test feedback document. As such, the feedback document was prepared in advance, but emailed to participants in early June, approximately two weeks after the final MCQ/simulation session was held.

The ideal timing for feedback depends on the task being assessed, the nature of the feedback, as well as characteristics of the recipient of the feedback. Feedback timing is generally classified as immediate or delayed. In general, feedback should be given soon after the assessment task is completed (i.e. immediate feedback), but should not interrupt the task. It is also important to allow the examinee to have some time to reflect on his/her own performance prior to giving feedback. This is particularly important if feedback is provided in an interactive form, with the assessor and examinee discussing the examinee’s performance together.

As such, the timing of feedback given for the MCQ test was not ideal, and detracted from the formative validity of the MCQ test. The delay between completing the assessment and receiving feedback was too long for the participants to necessarily recall which answers they chose and the thought processes that led them to those answers (especially since their actual exam papers were not returned to them). While participants may have initially been interested in receiving feedback on their performance and the correct answers (note that individual marks were never given, this is also discussed in detail), they would likely have lost interest by the time they received the feedback document, especially since this assessment did not contribute in any way to their marks.

The form of the feedback was a PDF document containing the MCQ items with
answers and explanations given for each item. Individual participant scores were not given, nor was any information given on the overall mean score or cohort mean scores. Email was used as the means for disseminating MCQ feedback document predominantly for practical reasons: most participants had supplied their email address when they expressed interest in participating in the study. Participants who had not supplied their email address were still sent the feedback document - for students via Northern Clinical School administration, and for junior doctors through the Northern Health staff email service. Given that both students and junior doctors are required to regularly check their emails, all participants should have received the feedback document in a timely manner. No issues with the method for disseminating feedback for the MCQ test were identified.

Written feedback is considered an appropriate, and even preferred, form of feedback for written assessments, such as MCQ tests. A study by Murdoch-Eaton & Sargeant suggested that medical students, particularly more junior students, place greater value on written feedback and perceive this as a more formal form of feedback. However, there is no specific form of feedback that is best for any type of assessment: the best form of feedback actually depends on the recipient, and different recipients will prefer or learn more from different types of feedback. According to Brinko, “feedback is more effective when the recipient is able to select the way in which it is conveyed.” Regardless of the form, experts recommend that all feedback should also be discussed with learners; this is discussed further in relation to feedback type.

In regards to the specific form of the MCQ feedback document, the MCQ items were in the same exact same form and order as presented for the MCQ test, and were
accompanied by answers and explanations for each item. One of the most important components in providing feedback for MCQ tests is ensuring that examinees are provided with the correct answers to the questions, so that they can learn from the assessment experience: this is really the essence of formative assessment.\(^{(11, 108, 109)}\)

This format was chosen so that participants could refer directly to the assessment items, recall their thought process in selection of an option (though, as previously mentioned, the delay in sending out the feedback document likely made this less likely), and then use the answers and explanations given to reflect on whether their thought processes, and final responses, were correct. This method allowed for important aims of feedback to be met: participants were able to identify gaps in their knowledge and flaws in their clinical reasoning, as well as acquiring new knowledge and guidance in clinical reasoning through the explanations, or even confirming that their knowledge and thought processes were correct.\(^{(108, 112)}\)

Participants were not given their scores, for either the MCQ test or the simulation assessment. The reasoning behind this was two-fold. Firstly, the actual scores achieved were not considered to be important given the context of the assessments- these assessments were undertaken for research purposes only. They did not contribute to any grades, were not criterion referenced in relation to university assessment, and were not reported to the university or hospital medical education unit. Regardless of the purpose of the assessment, in formative assessment, marks are less important than constructive feedback, so it was not thought that releasing marks to participants would be particularly useful.\(^{(112)}\)
The type of feedback provided was general/group, unidirectional feedback, with a focus on item stems/clinical context, rather than just the correct answers. Feedback is most useful when it is individualised and provides information on an individual’s performance and areas for improvement. (91, 100, 111) Group feedback (small- or large-group) is used, and still provides some information, but is less useful than individual feedback, since learners are individuals, and they therefore require different feedback. (100, 113, 114) The lack of individualised feedback provided for this MCQ test significantly detracts from the formative validity of this assessment tool. As previously discussed, formative validity, and the quality of formative assessment, relates to the usefulness of information provided to both the examinee/learner and the assessor/educator. (11, 13-16) In this case, while useful information is provided to the participants as a group, the feedback is not specific to any individual, such that the individuals must search through the feedback document and reflect on their own performance, in order to gain any useful information. Again, the reason for giving group feedback rather than individualised feedback, was related to a lack of resources and concerns regarding maintaining the integrity of the research assessment tools, as discussed above in relation to reasoning for not releasing participant scores.

The feedback provided for the MCQ test was uni-directional, that is, information was provided to participants but there was no opportunity for discussion between the participants and the assessor in regards to the feedback. Experts recommend that with every instance of feedback, there should be an opportunity for discussion, preferably face-to-face. (91, 92, 110, 114) This is also the basis of most feedback models used in medical education, such as Pendleton’s Rules, The Chicago Model, The SCOPME Model, The Six-Step Problem-Solving Model, SET-GO and ALOBA (agenda-led, outcome-based analysis). (11, 114) Unfortunately, logistical/resource issues meant that
it was not possible to discuss feedback for MCQs with individual participants. However, each participant did receive individualised feedback on their performance on the simulation assessments, with opportunity for discussion as part of the debrief.

In retrospect, the decision not to provide participants with the results of their MCQ assessment may have limited the formative validity of this assessment method. Given the timing between the participants completing the MCQ test and receiving the feedback document, they would likely not be able to reflect properly on which items they got wrong, as they may not have been able to remember which answer they selected, or the thought processes that led them to select a particular answer. In the setting of this research project, the ability to provide immediate feedback was limited by the importance of maintaining the integrity of the assessment instrument (i.e. participants not discussing answers). However, alternative methods of feedback could have been used, such as:

- Providing participants with a numerical result (either soon after they participated, or at the time when the feedback document was distributed): this is relatively low level feedback but still provides participants with more information on their performance than they would otherwise have obtained.
- Providing a forum for discussion of the questions and answers after the completion of the research project (e.g. a discussion session or tutorial). This would have significantly improved the formative value of the assessment tool.

Explanations that were given for the MCQ items provided teaching around the topics/learning outcome categories being assessed. The correct answers were given and justified, based on an explanation pathophysiology/pharmacology, or conceptually, or with reference to clinical practice guidelines. Importantly, the incorrect options were
also discussed, and justifications were given as to why these were not the best options.

Most explanations provided a general statement on the utility/applicability or interpretation of a given investigation or management strategy, such that the teachings could be applied to any case of the same pathology (AMI or APO). For example, the feedback provided for item 1 on the MCQ test is shown below:

**Segmental systolic dysfunction on transthoracic echocardiogram (TTE) suggests which underlying cause of heart failure?**

A. Hypertrophic cardiomyopathy  
B. Valvular disease  
C. Ischaemic heart disease  
[answer: C]

*Explanation: Segmental systolic dysfunction on TTE describes dysfunction localised to a specific area of myocardium. This suggests ischaemic heart disease, as a territory (or territories) supplied by a diseased coronary artery (or arteries) is affected but there is not global dysfunction. Hypertrophic cardiomyopathy appears on TTE as areas of hypertrophied myocardium and generally affects the left ventricle. Valvular disease is demonstrated by abnormal flow of blood through a valve (or valves), and can cause dilatation or hypertrophy of isolated chambers, but this is distinct from segmental dysfunction.*

In this case, the explanation gives the definition of the main term in the stem, *segmental systolic dysfunction*, discusses the implications of the term (*ischaemic heart disease*) and the basic pathophysiology (*coronary artery territory affected*), thereby justifying option C as the correct answer. Then the transthoracic echocardiography findings that would be associated with the alternative options is discussed, such that the examinees learn about common echocardiography findings of a few common pathologies, rather than only ischaemic heart disease.

Where the items related to specific vignettes/clinical situations, the explanations sometimes related back to the clinical context but then discussed how this could be related to a type/category of clinical situation or patient, to provide more generalisable learning. Given that the purpose of formative assessment is to guide further learning, the
intention in the construction of the MCQ test feedback document was to go beyond provision of the answers, and provide an opportunity for learning about commonly encountered, and commonly examined, aspects of AMI and APO management. The format of the feedback document (items and options given in full) also allows candidates to use the test as an aid to revision, i.e. they could use these as practice questions for exam preparation.

The quantity of feedback given was appropriate to the context of group feedback, because it was necessary to provide relatively detailed explanations for each item. If the feedback had been individualised, it may have been appropriate to identify, more briefly, the examinee’s strengths and weaknesses, giving examples. (11) Generally when individual feedback is given for an MCQ test, the focus is on the items that the examinee answered incorrectly. (108) However, it has been shown that it is also important to give examinees feedback on items that were answered correctly, for the purpose of “enhancing retention of low-confidence correct responses”. (108) It is important to acknowledge that correctly answering an MCQ item does not necessarily correlate to mastery of the concept being assessed by that item, and so feedback on items answered correctly but with low confidence is crucial. (108) Taking this into consideration, it could be argued that the form, type and quantity of feedback included in the MCQ test feedback document was actually the most appropriate/useful for this type of assessment. Supplementation of this feedback document with a short summary of the individual’s performance (test score, areas of strength and weakness) may have been useful.
All feedback provided was accurate, and based on sound sources of information. As discussed in Chapter 4: Assessment Tool Development, evidence-based resources and local guidelines were used in the construction of MCQ items and confirmation of correct answers. These resources were also employed, where required, in construction of the MCQ feedback document. Qualitative analysis of the document did not reveal any concerns about the accuracy or appropriateness of the explanations given.

References were only cited twice in the MCQ feedback document, where direct quotations were used. In retrospect, provision of references to examinees, either as a list at the end of the document, or related specifically to individual items, would be useful to promote further reading/learning. While the explanations themselves provide some teaching, some examinees may prefer to be guided to more comprehensive resources, whether to clarify concepts they found difficult, or out of interest.

iii. Formative Validity Summary

Overall, quantitative and qualitative analysis demonstrates that the MCQ test has good formative validity, within the limitations inherent to this assessment method. Unfortunately, MCQ tests do not allow the assessor to observe/scrutinise the examinees’ clinical reasoning processes or application to practice. This means that there is limited data from which the assessor can appraise the examinee. Therefore, MCQ test inherently have lesser formative validity than more practical forms of assessment.
The major strengths and weakness identified for this specific MCQ test, in relation to formative validity are summarised as follows:

- **Strengths:**
  - good spread of item difficulty levels- overall and across learning outcome categories
  - feedback given on every item
  - comprehensive explanations for each item, addressing both correct and incorrect answers
  - generalised teaching through explanations

- **Weaknesses:**
  - in general, poor ability of individual items to differentiate between candidates’ performance
  - group feedback
  - unidirectional feedback

In summary, the MCQ test, in conjunction with the MCQ test feedback document, satisfied the requirements for an assessment to have formative validity; it provided useful information for the assessor on individual and cohort performance, and provided useful feedback to direct further learning for examinees. (11, 13-16) In retrospect, though, provision of individual results to the participants would have improved the utility of this assessment as a formative assessment tool. Beyond this, the utility of the assessment, in terms of formative validity, depends on educators’ and learners’ ability to reflect on the information/data to guide further teaching and learning.
B. Simulation Assessments

i. Quantitative Findings

As discussed previously in relation to simulation assessment development, the simulation assessment checklist items correspond to necessary and desirable steps in assessment and management, and are not intentionally of differing difficulty, as they would be for a written assessment. However, those examinees with greater clinical proficiency would be expected to address a greater number of the items. Therefore, those items that would be addressed by the higher performing candidates but not the lower performing candidates represent the more ‘difficult’ items. These items are not inherently difficult because they require a greater level of knowledge or skill, but rather they may represent areas that are commonly overlooked, or that would differentiate between a basic assessment and a comprehensive assessment. For example, the most ‘difficult’ item on Simulation Scenario 1 (the item with the lowest mean score) was requesting a chest x-ray for a patient presenting with chest pain. The most ‘difficult’ item on Simulation Scenario 2, was obtaining history about whether the onset of dyspnoea was exertional or at rest. Both of these items represent steps that may be unintentionally excluded by a less experienced clinician, but which could assist in differentiating between differential diagnoses.

Interpretation of the difficulty of items on a simulation assessment, which is almost entirely scored based on a checklist, is more difficult than for an MCQ test (for these simulation assessments, 27 marks were allocated to the checklist, 3 marks were allocated to the GAS). Whereas an MCQ test has a clearly defined number of items for examinees to complete, and for each item the examinees could select the correct option by chance (the probability of which depends on the number of options given), in a simulation assessment the number and nature of checklist items being assessed is not
known to the examinee, and marks are attained by performing key actions, therefore cannot be scored by chance. The mean item values that represent different levels of difficulty for an MCQ assessment are well-established, based on probability/statistical principles, but this is not the case for simulation assessments, therefore it is more difficult to interpret the mean score/difficulty value.

For an MCQ test, there is a baseline mean item score that would be attained by chance, which depends on the number of alternative options from which the examinee must select the correct answer. Where there are three options given, a mean item score of 0.33 would be attained by chance. (Where there are four options given, a mean item score of 0.25 would be attained by chance, and so on.) Based on this principle, for an MCQ test with three options per item, a mean item score of 0.67 is considered to represent medium difficulty, since this value lies halfway between the chance of answering the item correctly by chance, and the chance of answering correctly based on knowledge of the answer. (98) No similar discussion of mean item scores for simulation checklists has been identified in a review of the literature. However, based on the same conceptual framework, it would seem logical that for a checklist, where there is zero chance of completing an item by chance, and the chance of completing the item based on knowledge is 1, a mean item score of 0.50 would represent an item of medium difficulty. As previously mentioned, though, items are not intentionally of varying difficulty, in that they simply represent the steps that would be followed for optimal assessment and management. As such, mean item scores for simulation assessments are better interpreted as a reflection of candidate performance, rather than a property of the items, and therefore will be referred to as mean item values, rather than item difficulty values.
Simulation Scenario 1 mean item values ranged from 0.32-1.00, though fifteen of the 26 items had mean values in the range 0.81-1. Simulation Scenario 2 had a broader spread of mean item values, with a full range 0.00-1.00. Items were more evenly spread across the range for Scenario 2. This difference in spread of mean item value for the two scenarios, apart from being reflective of Scenario 2 being more difficult, is also indicative of the greater ability of Scenario 2 to differentiate between candidates across the full range of abilities. The presence of items with a range of mean item values also suggests that the simulation scenarios are of sufficient difficulty to allow the examiner/educator to assess examinees’ level of performance for each component, thereby improving the formative feedback. If all items are addressed by all examinees then this would suggest that the scenario is of insufficient complexity to differentiate between examinees, and therefore is not useful for providing formative feedback.

Tables 14.9 and 14.14 show the range of mean item values by learning outcome category. For all learning outcome categories that were assessed by more than one item, for both simulation scenarios, there was a range of mean item values, thereby allowing differentiation of candidates’ performance in each of these learning outcome categories. For Simulation Scenario 1, there was a very even spread of mean item values for the history component, and a reasonable spread for other aspects of patient assessment. However, mean item values were mostly higher for the patient management learning outcome categories, suggesting that examinees better addressed aspects of management, compared to aspects of assessment. For Simulation Scenario 2, mean item values for items corresponding to history were lower than for Simulation Scenario 1, suggesting that examinees are less able to take a comprehensive history for dyspnoea, compared to chest pain. Examination was approximately equally well-performed for dyspnoea and chest pain, which is expected given that the required general cardiorespiratory
examination was the same for both scenarios. Investigation was less well-addressed for Simulation Scenario 2. Overall, for Simulation Scenario 2, management items had higher mean scores than assessment items. However, mean items scores for management components were slightly lower than for Simulation Scenario 2, suggesting less overall competence in managing APO than AMI.

The results of the Mann-Whitney U test, shown in Tables 14.10 and 14.15, indicate that there were two items on Simulation Scenario 1 (items A1c and A4a), and two items on Simulation Scenario 2 (items A4a and A4c), where there was a statistically significant difference in performance between ‘novice’ and ‘expert’ examinees. For item A1c on Simulation Scenario 1, the ‘novice’ (MD3) group performed better than the ‘expert’ (HMO2) group, and this difference in performance was statistically significant. This is an unexpected finding, so it is important to examine this item. This item required examinees to enquire as to the character of chest pain. It is surprising that HMO2 doctors would not obtain this part of the history, given that the character of pain is one of the properties used to differentiate typical ischaemic chest pain from atypical chest pain. (115) A possible explanation for ‘novice’ examinees more reliably addressing this criterion is that medical students are more likely to approach a chest pain history with a set of learnt questions, such as those included in the commonly taught mnemonic SOCRATES (site, onset, character, radiation, alleviating factors, time course, exacerbating factors, severity). (115) Another consideration that may have caused the ‘expert’ group to neglect this question, is prioritisation of tasks; that is, recognising chest pain as a presentation requiring urgent assessment and management, more experienced examinees may have prioritised obtaining and ECG and giving first line treatment, over taking a thorough history. This difference in the approach to competing tasks is also discussed in Chapter 22: Qualitative Analysis of the Research Process.
For the remainder of the items identified (Scenario 1 item A4a, Scenario 2 items A4a and A4c), as expected, the ‘expert’ group performed significantly better than the ‘novice’ group. As previously discussed, identification of such key items could allow educators to use performance on these specific items to gauge overall performance. It is also important to determine the content being assessed by these items, to identify areas that may require more teaching in order for ‘novices’ to eventually progress to mastery of AMI and APO management.

As explained previously, given the multiple comparisons being performed in assessing the items in each simulation assessment tool, it would be appropriate to apply a Bonferroni correction in interpretation of the results.(83, 105, 106) The mathematical formula for calculating the adjusted significance level is included in Chapter 19: Construct Validity, and depends on the number of comparisons being made (the number of assessment items). The calculated adjusted significance levels both for the simulation assessments is \( \alpha = 0.0019 \), i.e. a p-value of less than 0.0019 is required to reject the null hypothesis.

This means that if the results of the Mann-Whitney U test are adjusted for multiplicity, there are no items on Simulation Scenario 1 for which there is a statistically significant difference between ‘novice’ and ‘expert’ performance. However, both items identified by the Mann Whitney U test for Simulation Scenario 2 (items A4a and A4c), remain statistically significant when corrected for multiplicity.

The results of the Kruskal-Wallis test, presented in Tables 14.11 and 14.16, indicate that Simulation Scenario 1 items A1c and A5d, and Simulation Scenario 2 items A4a and A4c, demonstrated a significant difference in performance between participants across
the spectrum of classifications. The results of significant pairwise comparisons for each classification are also presented. It is important to note that these results include a Dunn-Bonferroni correction for multiplicity. That is, the p value itself is corrected for multiplicity and quoted as an adjusted significance.

Therefore, the important findings from the Kruskal-Wallis tests, after pairwise comparisons and Dunn-Bonferonni corrections, are:

- **Scenario 1, Item A1**: There was no significant difference between any two groups.
- **Scenario 1, Item A5c**: There were statistically significant differences in performance between MD2 and MD3 groups, and MD2 and HMO2 groups, with differences in the expected direction (more senior classifications performing better than more junior classifications). However, there was only one participant at MD2 level who was allocated to Simulation Scenario 1, therefore, there is no true mean value for this level, so this statistical analysis was not valid.
- **Scenario 2, Item A4a**: There was no significant difference between any two groups.
- **Scenario 2, Item A4b**: There was a statistically significant difference in performance between MD3 and HMO2 groups, with differences in the expected direction. However, performance did not continuously improve across the range of classifications.
- **Scenario 2, Item A4c**: There were statistically significant differences in performance between MD2 and Intern groups, and MD2 and HMO2 groups, with differences in the expected direction. There was continuous improvement in performance across most of the classifications (MD2-Intern), suggesting ability to discriminate across a range of levels based on this item.
• Scenario 2, Item A4d: there was a statistically significant difference in performance between MD2 and Intern groups, with differences in the expected direction. However, performance did not continuously improve across the range of classifications.

As previously discussed, small sample sizes make interpretation of results for subgroups of the entire study population difficult. This is discussed in more detail in Chapter 25: Limitations.

ii. Qualitative Findings

Figure 14.5 shows the results of the qualitative analysis of the simulation assessments, listed under headings that describe the important aspects of feedback: form, type, quantity and quality.\(^{(91, 92)}\)

Feedback was verbal, which is appropriate to the assessment type.\(^{(91)}\) Verbal feedback is also the standard method for debriefing after a simulation experience, and facilitates feedback being immediate and discussion-based.\(^{(41)}\) However, some experts assert that feedback should be in written form, and that written feedback is perceived by learners to be more valuable.\(^{(92, 111)}\) While this may be appropriate for more formal, summative assessment, it may not be as necessary in the case of formative assessment. In fact, all of the commonly used feedback models in medical education, including Pendleton’s Rules, The Chicago Model and The SCOPME Model, among others, use verbal feedback.\(^{(11)}\)

Timing of feedback in relation to the assessment is important and, as discussed earlier in this chapter, the optimal timing for feedback depends on the task being assessed, the
nature of the feedback, as well as characteristics of the recipient of the feedback. (91, 100, 108) In this case, feedback was provided immediately after the conclusion of the ten-minute simulation. This meant that participants were able to reflect on their performance while they were still able to remember the details, but the feedback did not interrupt their performance of the assessment task. (11, 91, 110) It was felt that this was the most useful and convenient timing for feedback, as it would be more difficult to organise for participants to return at a later date/time for feedback, and both participants and assessors would be less likely to remember details of the participants’ performance later (even if assessors could refer to the assessment sheets), and therefore the feedback would be less useful.

The structure of the debrief was included in Simulation Scenarios document (Appendix 5: Simulation Scenarios). The structure of the debrief is based on Pendleton’s Rules, (11) with an additional two points allowing for some teaching on the topic of the simulation (AMI or APO). As such, the structure of the debrief was based on sound educational principles, and would have enhanced the formative validity of the simulation assessment tools.

In regards to the type of feedback provided in the debrief process, feedback was individualised and involved discussion between the participant and assessor. Both these aspects of the debrief indicate a high level of formative validity, as the quality of formative assessment relates to the usefulness of information provided to both the examinee/learner and the assessor/educator, and the approach used allows for maximal transfer of feedback between these two parties. (11, 13-16) Feedback is most useful when it is individualised and provides information on an individual’s performance and areas for improvement. (91, 100, 111) Experts also recommend that all feedback should
also be discussed with learners, preferably in the form of face-to-face discussion between examiner and examinee. (91, 92, 110, 114)

Reflection is another important part of feedback and learning, particularly for adult learners. (11, 14, 114) The debriefing structure used for the simulation assessments encourages reflection by asking participants to comment on, and critique, their own performance. The structure also required both the assessor and participant to identify both positive and negative aspects of the participant’s performance, which is a necessary characteristic of good quality constructive feedback. (11)

As previously mentioned, the modification that was made to Pendleton’s Rules for the debrief structure was the addition of teaching points. This was important in the context of an assessment that was used for research purposes, rather than as part of a teaching programme. Usually the information that the educator obtains from the assessment, which is an important but sometimes forgotten aspect of formative assessment, is used to guide further teaching. (16, 51) Since this would not be possible in this context, it was important that there be some immediate teaching about the topic of the simulation (AMI or APO), in cases where it was felt that participants would benefit from this. Some participants had clearly understood the clinical scenario and performed very well, but may have simply forgotten a small step (e.g. an isolated part of the history, examination or investigation) and did not require any teaching beyond the feedback (steps 1-4 of the structured debrief). Other, more junior, participants sometimes required more direct teaching, especially on how to manage AMI or APO.

The quantity of feedback was appropriate for the assessment. With five minutes allocated to debrief after a ten-minute simulation, there was more than adequate time to
discuss the participant’s performance and provide some teaching on the topic. Some participants required less than the allocated time; and for those participants who could benefit from it, and were keen to be taught, the debrief time was extended beyond five minutes to allow for more teaching.

The feedback provided was relevant, specific to hospital protocols/practice, and delivered by assessors with appropriate expertise. The assessors used all had specific cardiology experience (i.e. had worked in a cardiology department), with the exception of assessor 4, who was a HMO2 level doctor, who was required to fill in for another assessor, and only acted as the assessor for two participants. Assessors 1 and 5 were HMO2 level doctors with experience as cardiology resident medical officers. Assessor 2 was a general medical advanced trainee (registrar) and assessor 3 was a cardiology research fellow and general medical advanced trainee (registrar). As such, all assessors had adequate cardiology knowledge and experience to give feedback and deliver teaching on AMI and APO topics.

Some assessors were less experienced than others and in some cases, the assessors were of the same level as some of the participants (HMO2). While it is important that those receiving feedback perceive those offering feedback as credible and knowledgeable, some studies have also shown that “feedback is more effective when the source of feedback is lower or equal in status to the feedback recipient”, and that “problems may arise when the source of feedback is higher in status than the recipient”. (110) A study by Cooper et al also demonstrated that emergency medicine residents are perceived, by medical students, to be equally effective in simulation debriefing, compared to emergency medicine faculty. (41) It is possible that HMO2 participants may have felt embarrassed about being observed and assessed by their HMO2 peer assessors,
however, this reflects real clinical practice, where it is not uncommon for peers to be present and observe each other’s practice, especially in an emergency situation. As such, the assessors used were of an appropriate, intermediate level relative to the participants.

While only two of the assessors had previous education in delivering feedback, the structured debrief format helped to ensure that the feedback received was of high quality, and of use to the participants.(11) The observers were not instructed on what specific information to provide in the teaching component of the debrief, and were allowed to teach as they saw fit. All information given, however, was based on clinical experience and Australian guidelines. More specifically, information was based on hospital-specific practices and protocols, so that it could be applied in practice. For example, a lot of feedback was given on how to seek senior assistance when required, and this included giving information on how and when to use emergency codes, specifically MET (Medical Emergency Team) Call, and Code STEMI. Indications for urgent cardiac catheterisation were also discussed, which again are, to some degree, institution-dependent. This is important practical information for participants, which might not necessarily be provided in classroom-based teaching.

It is important to consider that, while the form, type, quantity and quality of feedback is important, and these were the categories used to evaluate formative validity, there are other aspects of feedback that also influence formative validity. In particular, feedback recipient (participant) factors very heavily influence on how well feedback is received and used. These participant factors include locus of control, self-esteem, amount of experience, developmental stage, preferred feedback method and engagement with the
Therefore, regardless of the characteristics of the feedback, the learning taken away from the debrief process is largely dependent on the individual.

While quantitative analysis provides useful formative information for the educator, as it evaluates the performance of the whole class, thereby providing guidance for further teaching, there is useful information to be gleaned from the qualitative data also. For example, there were several observations made about the differences in performance with experience (both approach and participant scores), and the shortfalls of the use of a predominantly checklist-based assessment tool. These are discussed in some detail in Chapter 22: Qualitative Analysis of the Research Process.

The feedback provided to participants did not specifically address the approach taken by participants (i.e. use of a systematic approach or a more integrative/parallel approach). However, the relatively open structure of feedback used, which encouraged discussion on any point relevant to the participant’s performance (rather than just giving scores and listing points missed), lends itself to discussion on the participant’s approach to assessment and management, as well as any incorrect assumptions or management decisions made by the participant.
iii. Formative Validity Summary

Overall, quantitative and qualitative analysis demonstrates that the simulation assessments have very high formative validity. The major strengths and weaknesses identified for the simulation assessments, in relation to formative validity are summarised as follows:

- **Strengths:**
  - individual feedback
  - immediate feedback
  - discussion-based feedback, including reflection
  - opportunity for some teaching at the time of debrief
  - adequate time for feedback
  - provision of relevant feedback
  - identification of ways in which participants of different levels approach the same clinical scenario
  - very good spread of mean item scores- overall and across learning outcome categories- allowing for differentiation of participants’ abilities

- **Weaknesses:**
  - no written feedback
  - in general, poor ability of individual items to differentiate between candidates’ performance

In summary, the simulation assessments had very high formative validity. They provided useful information for the assessor on individual and cohort performance, and provided useful feedback to examinees to gauge their performance (without comparison to other participants) and direct further learning.(11, 13-16)
C. Comparison of the Multiple Choice Question Test and Simulation Assessments

Quantitative analysis of the formative validity of both the MCQ test and simulation assessments found that there was a good spread of mean item scores/difficulty levels, overall and across the different learning outcome categories. As previously discussed, there was a different focus in interpreting mean item scores for the two different types of assessments; for MCQ tests, the mean item scores were predominantly interpreted as representative of item difficulty, since there was an intention to include items at a range of difficulty levels. For simulation assessments, this value was considered to be more of an indication of examinees’ performance, rather than representing difficulty. Regardless, a range of mean item scores means that the assessor/educator is provided with adequate information to evaluate each individual’s performance, regardless of their level, across the range of learning outcome categories, in order to determine areas for improvement.

Quantitative analysis demonstrated that, despite the ability of the MCQ assessment tool and Simulation Scenario 2 to differentiate between candidates of ‘novice’ and ‘expert’ level at the level of performance on the entire assessment (thereby demonstrating construct validity), individual items on these two assessments were generally unable to differentiate between candidates of different levels. This was also the case for items on Simulation Scenario 1, though this was perhaps more expected, given that this assessment tool, as a whole, was unable to differentiate between ‘novice’ and ‘expert’ examinees.

Qualitative analysis of the MCQ test feedback document showed that adequate information was provided to students and educators to satisfy requirements for
formative feedback. That is, the information obtained from the assessment was able to guide further teaching and learning. However, consistent with the inherent limitations of MCQ tests in providing useful information on examinees’ clinical reasoning processes or application to practice, the MCQ test had much lower formative validity than the simulation assessments. The formative validity of the MCQ test was also reduced by the timing of the receipt of feedback, which was too delayed to allow for optimal learning. This is in contrast to the simulation debrief, which provided immediate feedback.

The format of the simulation assessment provided assessors with much more useful information to guide further teaching, by allowing observation of clinical performance. This would be of great benefit if the simulation assessments were used in the context of an educational program (rather than a research context). In addition, the simulation debriefing provided an opportunity for examiners and examinees to discuss performance on the assessment, and participants were able to receive immediate, individualised, relevant feedback, as well as some teaching. Participants were also able to ask questions or clarify areas of concern/confusion. This is in contrast to the written group feedback that was provided for the MCQ test, which had lower formative validity because it was unidirectional and was not customised to the individual participants’ areas of weakness/gaps in knowledge.

As briefly mentioned in Chapter 18: Content Validity, when choosing an assessment method, it is also important to consider feasibility and, therefore, compare the efficiency of the assessment types. With regards to formative validity and the amount of information obtained for use in guiding further teaching/learning, the MCQ test assessed both topics using a total of 30 items, whereas the simulation assessments assessed each topic with just less than 30 items. Therefore, the simulation assessments
were able to provide more information about the candidate’s performance in the same amount of testing time. While this superior testing efficiency on an individual level is important,(59) in considering feasibility this would need to be balanced against the greater resources required for the simulation assessments (which need to be performed individually) compared to the MCQ test (which all participants can complete simultaneously).

This chapter has presented the research findings on Research Question 1B- “What is the formative validity?” and has contributed to answering Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?” The next chapter will discuss the research findings related to Research Question 1D- “What is the inter-observer reliability of cardiac emergency simulation as an assessment tool?”
CHAPTER 21: INTEROBSERVER RELIABILITY

This chapter will address Research Question 1D- “What is the inter-observer reliability of cardiac emergency simulation as an assessment tool?”

As previously discussed, in any assessment where the assessor must make a judgement on the candidate’s performance, such as assessments that use checklists, there is propensity for subjectivity in scoring. This may be due to poorly defined/subjective criteria or differences in examiners’ expectations.\(51, 58\) It is important to evaluate the degree and source of subjectivity, or interobserver reliability, of any assessment tool, in order to be able to minimise any discrepancies.\(51\)

Percentage agreement between examiners for overall scores and component scores was variable (Table 15.1). In general, there was better percentage agreement between assessors on GAS score, compared to checklist score. This is not unexpected because, despite GAS potentially being more subjective (even with criterion-based allocation of marks for the GAS component), the possible range of scores for GAS is much smaller (0-3) than the checklist component (0-27). Assessors would be expected to differ less across a smaller scale, as there is less scope for variance.

Focusing on agreement for the overall score (since this incorporates both checklist and GAS components), the combination of assessors 1 and 5 had the highest percentage agreement, at 50% (for each simulation scenario individually, and both scenarios combined). This was much higher than the other combinations of assessors.

There is no consensus as to the percentage agreement that is acceptable or desirable for interobserver reliability, because the interpretation of the percentage agreement depends
on the context and, specifically, the likelihood of agreement by chance.(94-96)

Therefore, it is instructive to consider the probability of agreement by chance for the measures we are analysing, in order to put into perspective the percentage agreement achieved.

**For each individual checklist item:**

For all items except for Scenario 1 item A3a, scoring is binary, i.e. there are two possible scores: 0 or 1. Therefore there is a 1 in 2 (50%) probability that both assessors would select the same score by chance.

For Scenario 1, Item A3a, scoring is from 0 to 2, therefore there are three possible scores, and a 1 in 3 (33%) probability that both assessors would select the same score for this item by chance.

**For the checklist score:**

Scoring is from 0 to 27. Therefore there are 28 possible scores, and a 1 in 28 (3.6%) probability that both assessors would select the same score for the checklist by chance.

**For the GAS:**

Scoring is from 0-3. Therefore there are four possible scores, and a 1 in 4 (25%) probability that both assessors would select the same GAS score by chance.

**For the overall score:**

Scoring is from 0-30. Therefore there are 31 possible scores, and a 1 in 31 (3.2%) probability that both assessors would select the same overall score by chance.
With this perspective, it is apparent that the percentage agreement results obtained for both component scores and overall scores were very much higher than could be achieved by chance. As such, using percentage agreement as a measure, interobserver reliability is high.

Spearman correlation has the advantage of considering rank, rather than absolute scores, when comparing assessors’ scores. (87) This would then account for assessors with particularly lenient or stringent scoring, because it is the ranking of examinees, rather than the absolute scores, that are important. (116)

Correlation coefficient values can be interpreted as follows: (117)

\[
\begin{align*}
\rho < 0.20 & \quad \text{No correlation} \\
0.20 < \rho < 0.40 & \quad \text{Weak correlation} \\
0.40 < \rho < 0.60 & \quad \text{Moderate correlation} \\
0.60 < \rho < 0.80 & \quad \text{Strong correlation} \\
\rho > 0.80 & \quad \text{Very strong correlation}
\end{align*}
\]

Spearman correlation coefficients (\(\rho\)) were very high and statistically significant for Scenario 1 checklist and overall score. GAS correlation scores for Scenario 1 were lower and not statistically significant, though there was still moderate correlation between ranks for assessors 1 and 5, and strong correlation between ranks for assessors 1 and 3. For Scenario 2, correlation was lower than for Scenario 1. For assessors 1 and 3, the correlation for Scenario 2 was strong but not statistically significant for overall mark and strong and significant for checklist score, but only moderate for GAS. For assessors 1 and 5, there was strong but not statistically significant correlation for overall mark and checklist score, but weak correlation for GAS.
Overall, then, Spearman correlation showed that there is very strong correlation between assessors in overall scoring for Simulation Scenario 1, and strong correlation between assessors in overall scoring for Simulation Scenario 2. This indicates high interrater reliability according to this measure.

Cohen’s kappa was variable between different combinations of assessors, and for the different components. Interpretation of kappa values, compared to percentage agreement, is more straightforward. Values are interpreted as follows:

\[ \kappa < 0.40 \] poor agreement

\[ 0.40 < \kappa < 0.60 \] fair agreement

\[ 0.60 < \kappa < 0.75 \] good agreement

\[ \kappa > 0.75 \] excellent agreement

Therefore, at best (for the combination of assessors 1 and 5), interobserver reliability as measured by Cohen’s kappa for both simulation assessments combined, and for Simulation Scenario 1, is fair and statistically significant and for Simulation Scenario 2 is poor and not statistically significant. The only limitation that must be considered in interpreting the kappa values obtained, is that where there are less than ten cases, accuracy may be compromised. Therefore, the accuracy of the results is only guaranteed for both simulation scenarios considered combined, for the combinations of assessors 1 and 3, and assessors 1 and 5 (not assessors 1 and 2).

Individual checklist items were analysed for percentage agreement to determine if there were particular items where assessors were more likely to concur or differ in their scoring. Scenario 1 had ten items (out of 26) with overall percentage agreement of 100% and Scenario 2 had 13 items (out of 27) with overall percentage agreement of 100%. The lowest overall (all assessor) percentage agreement was for Scenario 2 item
A1a, which had an overall percentage agreement of 35.3% (for all assessors), and a percentage agreement of 0% between assessors 1 and 5. Other items with low overall percentage agreement were Scenario 2 item B2a (70.6%) and Scenario 2 item B2b (58.8%). All other items had an overall percentage agreement above 75%.

Item A1a assessed whether or not the participant asked about onset of dyspnoea. A possible explanation for the discord between examiners on this point may relate to the ambiguous wording of the criterion—this item was intended to ascertain if onset was sudden or gradual, but could have been interpreted as asking about the time of onset. Scenario 2 item B2a assessed whether the candidate re-evaluated the patient’s dyspnoea, and item B2b assessed re-evaluation of vital signs. For these items, it is more likely that the issue was not with the wording of the criteria, but rather difficulty ascertaining whether the candidate had adequately re-evaluated the patient. Based on observations and qualitative data (field notes), it was sometimes difficult to determine if the participants were observing vital signs on the bedside monitor, especially later in the simulation, when they may have just glanced quickly over at the monitor to look for improvement. It was also difficult to determine if the participant had adequately re-evaluated for dyspnoea, as many candidates simply enquired as to whether the patient was “feeling better”. Whether the assessor considered this adequate re-evaluation of dyspnoea was subjective. These were the most subjective of the checklist items, as the other items were more easily assessed as either completed or not completed.

Overall, there were few items where there were systematic issues with disagreement between examiners. Bearing in mind the issues with the three items identified above, some modifications could be made to these items to improve interobserver reliability.
The contribution of observer error to inter-observer differences and measures of inter-
observer reliability must also be considered. Some minor variation in results would
likely be attributable to observer error. In a checklist-based assessment in particular, if
one examiner fails to observe completion of an action/task (e.g. due to distraction or
obstructed view) but the other examiner does observe it, then unless the examiners are
permitted to discuss their marking, variation in scores will occur. Markers were not
specifically told not to discuss their marks in this case, and in a few cases the examiners
did discuss whether or not a particular checklist task was observed. However, in
general, discussion between examiners before completion of the assessment was
discouraged.

Another factor potentially contributing to observer error in these simulation assessments
was failing to assign a yes/no mark to each checklist item. In many cases, examiners
simply marked those criteria that were completed (i.e. circled/ticked yes) but left those
that were not completed blank (presumably planning to come back to those items if they
were completed later). Those criteria that were not marked in any way were assumed to
have not been completed by the candidate and were scored as such (zero mark).
However, it is possible that the examiner erroneously skipped these criteria. Each of
these factors would have contributed to observer error, and therefore have had an
impact on interobserver reliability.

Across all measures of interobserver reliability, assessors 1 and 5 had the highest
concordance in scoring. A possible hypothesis to explain differences in interobserver
reliability for different combinations of assessors is that with more practice, marking
becomes more accurate, and therefore is more likely to agree with the other assessor.
This phenomenon of improved standardisation of scoring with practice and training has
been proven, (116) but it is difficult to determine how this affects scoring when there are multiple observers assessing different numbers of examinees and one constant observer.

It is important to note that assessors did not receive formal training in use of the assessment tools (e.g. did not have the opportunity to practice use of the tool on a standardised examinee or similar) (116), however, the tool was explained to assessors and there was the opportunity to ask questions/clarify any issues.

Assessor 1 had more experience in assessing these scenarios than all other assessors, since assessor 1 was constant for all participants. Assessor 2 initially had the same amount of experience as assessor 1, since this combination of assessors (assessors 1 and 2) scored the first five participants, however, both assessors would have been inexperienced in scoring their first five cases. If the agreement depended mostly on the variable assessor (rather than the constant assessor, assessor 1), then it would be expected that the combination of assessors 1 and 3 would have the highest agreement, since this combination of assessors was used for the largest number of cases, and assessor 3 had the most experience of all the variable assessors. However, this was not the case. If it was the influence of the experience of the constant assessor (assessor 1) that affected the interobserver reliability, then the interobserver reliability would be highest for the combination of assessors 1 and 5, since this combination marked the last participants. This, in fact, was the results reflected. It is possible, then, that assessor 1 improved in consistency with experience, and by one assessor marking more consistently, this resulted in overall improved agreement between the assessors.

Overall, combining all measures of interobserver reliability, there was very high interobserver reliability for Simulation Scenario 1 and fair interobserver reliability for Simulation Scenario 2. There was better agreement for the checklist component than for
the GAS; this is somewhat unexpected but may be attributable to the greater degree of subjectivity associated with the GAS, despite the GAS being criterion-based and having a smaller range/scale. This could suggest that the criteria of the GAS are lacking in clarity, or that there was an issue with overall assessor leniency/stringency. The interobserver reliability of the assessment tools, especially Scenario 2, could be improved by addressing and adding clarity to the few items that were identified as having lower interobserver reliability. Interobserver reliability could also be improved by providing training to assessors. (116)

Other factors that could have affected interobserver variability but were not specifically examined were halo effect and drift. (85, 86) Halo effect describes the phenomenon in which the assessor’s overall impression of an examinee (the term is usually used in relation to a positive impression) impacts their scoring of the examinee; for example, if an examinee presents particularly impressively at the commencement of an assessment, an assessor may tend to mark the examinee more favourably throughout the assessment. Drift describes an effect whereby an individual marker’s scores tend to increase or decrease over time. (86)

Halo effect would apply to both examiners for each simulation scenario. However, this would likely only affect the GAS component of the assessment, since there is little room for subjectivity or overall impression of the participant to affect checklist marking. Having criterion-based GAS scores would also decrease the subjectivity and potential for halo effect for the GAS component.
Drift would arguably only have significant implications for examiner 1, who examined all participants, since the other examiners marked a much smaller proportion of examinees; however, there was no evidence of any significant drift observed. Drift was likely minimised by a few factors:

- the fact that most of the total score was formed by the checklist component
  (less subjectivity)
- alternation between scenarios
- examinees of significantly different experience levels
  (effectively resetting the assessors’ expectations between simulations)
- having the simulation assessments spread over different days

This chapter has addressed Research Question 1D- "What is the inter-observer reliability of cardiac emergency simulation as an assessment tool?" The next chapter will go on to present findings in relation to Research Question 4- "What are the strengths, weaknesses, limitations and issues related to conducting simulation assessment research?"
CHAPTER 22: QUALITATIVE ANALYSIS OF THE RESEARCH PROCESS

A. Participant Recruitment and Logistics

Domains and dimensions were identified in relation to the topic of participant recruitment and logistics (Figure 22.1).

Time

The first domain identified was time. As discussed, the timeline of the research process was impacted by multiple external factors, and any alterations to the planned timeline had flow-on effects for subsequent steps in the research process. There was also a finite timeframe for the completion of the actual assessment component (MCQ test and simulation execution) of the research, which was limited by a range of factors.

Firstly, the simulation laboratory requires booking months in advance. In selecting the booking times, the number of participants able to be assessed within a session time was calculated, and it was determined that with a series of seven sessions spread over two weeks, it would be possible to accommodate a maximum of 128 participants. This was based on running MCQ and simulation assessments simultaneously, and simulation assessments consecutively, with only the time required for pack up/set up between participants. When it was determined that the actual number of participants that would likely be recruited was significantly less than the target, and very much less than the maximum able to be accommodated, it was decided that, in order to make transitions between the two scenarios easier, participants’ scheduled times would be spaced out more, allowing more time to transition between consecutive simulations.

Another limiting factor was the availability of participants- in many cases, participants
were only available towards the end of each session, such that there was difficulty
filling the earlier time slots, and most participants were scheduled at the end of the day.
Then, with the addition of some participants being recruited on the same day as when
they completed the assessments (as previously discussed data collection phases
occurred simultaneously), the last approximately two hours of the sessions involved
multiple participants completing the MCQ tests at the same time, and waiting for a short
time to complete the simulation assessments back-to-back.

Resources

The second domain that was identified as influencing participant recruitment was
resources. One of the most commonly cited concerns regarding the feasibility of
simulation in health professional education, whether as a teaching or assessment
method, is that it is a resource-intensive approach. (38, 66, 118-122) This is consistent
with the findings of this study, which identified five dimensions related to the resources
domain: simulation laboratory, assessors, actors, resource availability and logistics. All
of these dimensions were identified as factors that required considerable organisation
and had the potential to become barriers to running simulation sessions.

Northern Health has excellent simulation facilities, so there were no concerns about lack
of physical resources to conduct the simulations. Fortunate timing resulted in the
simulation laboratory having greater availability than usual over the period of time that
the research project was being conducted; however, the simulation laboratory is booked
at least weeks in advance, such that there was no ability to make alterations to session
times closer to the date.
Another difficulty that was encountered was in enlisting assessors for the simulation assessments. While there were clinicians and educators who were enthusiastic about the project, few of these were able to spare time to assist over at least half a day. If examiners are clinicians there is often some difficulty finding time away from clinical duties. During the simulation sessions, there was a situation where one of the examiners had to leave to attend to clinical duties, and it was not possible to find a replacement examiner to cover for the entire remainder of that scheduled session (a replacement examiner was available to assess two participants only). As such, there were two participants who were only assessed by one examiner for this reason. Again, the issue of finding available, suitably qualified assessors is not unique to this research project. Since this is a recognised potential barrier to use of simulation in health professional education, a small amount of research has been conducted to determine whether less experienced clinicians or even laypersons are able to fulfil the role of assessor.

The actors used for this research project were volunteers and were not paid for their time, since this was an unfunded research project. The actors who played the part of patient were not qualified actors, and had varying degrees of knowledge/experience in the healthcare field. The actors who played the part of nurse were required to be qualified nurses in order to be able to realistically imitate the nurse role. The impact of using untrained (volunteer) actors, in terms of affecting fidelity, reproducibility and consistency, is discussed in detail in the next section of this thesis. Given that the actors were not paid, and qualified nurses were required, it was difficult to find enough actors to run the simulations. This was a major hurdle, and almost prevented some sessions going ahead. For some sessions, the student researcher had to play the role of nurse as well as assessing.
The cost and availability of trained actors would be expected to be less of an issue for simulations run by medical schools or hospital medical education units, as they may be able to fund paid actors or possibly have access to a list of volunteer actors. If paid actors were used then the cost of roll out of simulation-based assessments would need to be considered when evaluating the feasibility of simulation as an assessment method.

Resource availability and logistics are interrelated, and in Figure 22.1, are shown in the overlapping area between time and resources. This is because, even if there are adequate resources available, it is very difficult to draw together the combination of resources that are required at the time when they are required.

**Interest**

The third domain identified was *interest*, with dimension of opportunity and learning. Given that participation was completely voluntary, the choice of whether or not to participate depended on potential participants’ interest in the project. Based on informal feedback, those who chose to participate did so because they felt that participation offered an opportunity that was different to those they were usually exposed to. Many participants reported that they had never had the opportunity to undertake a simulation individually, and to receive individual feedback. All medical students, and many junior doctors, said that they had not ever had the opportunity to run a cardiac emergency situation and make management decisions, and expressed that they appreciated the opportunity to do so in a safe, simulated environment. Other participants chose to participate because they thought that the experience would offer a good learning experience, especially given the feedback provided in the simulation debrief. Some junior doctor participants expressed that they felt that they received minimal teaching, and therefore took this opportunity to learn about an important topic.
Of course, if simulation was used as part of summative (or formative) assessment in the medical school curriculum or in junior doctor medical education, then this would circumvent the issue of poor recruitment. Participant interest would be much less important if participation was compulsory (though still relevant to assessment acceptability). It would also be easier to navigate the issues related to logistics if participants were required to attend for assessment at a particular time, with this assessment taking priority over other (clinical and educational) tasks or requirements, thereby removing a major barrier to participation.

**Communication**

The final domain identified in relation to participant recruitment and logistics was communication. The only official communication that potential participants received was the recruitment material that had been pre-approved by the ethics committees, and which is included in Appendix 6: Recruitment Material. This material was designed to give potential participants an overview of the purpose of the research, what would be required (in terms of participation and time commitment), and the potential benefit to the participants. As discussed, approved flyers with attached sign up sheets were placed around the Northern Hospital at Northern Clinical School, the education precinct, resident quarters and doctors’ write-up areas. Word of mouth is included in the overlap between interest and communication domains, because where there was interest on behalf of participants, or they found the experience of participating to be particularly useful, they tended to encourage others to participate.

The dimension competing priorities was identified as overlapping three of four major domains (time, interest, communication). The feedback that was received, both from participants, and from those who declined to participate, was that it is difficult to find
time to participate in a research project as they had many competing priorities, and they had to determine whether or not participating would be the best use of their time. The two factors that differentiated those who participated from those who did not, were: firstly, their perception of the recruitment material and any influence of word-of-mouth from peers/coworkers; and secondly, perhaps influenced by the former, their level of interest, and whether or not they perceived participation as a potentially valuable opportunity or learning experience.

It is evident that there were many factors that influenced participant recruitment and the logistical aspects of carrying out this research project. The significance of these findings is twofold: firstly, they contribute to our understanding of the research process; secondly, they help to explain the barriers that resulted in lower than target recruitment. By identifying these barriers, it is then possible to consider ways to address these barriers for future research. This is further discussed in *Chapter 25: Limitations*, and *Chapter 26: Future Directions*.

**B. Simulation Assessment Execution**

Qualitative analysis revealed five main domains in relation to the topic of *simulation assessment execution*: design, execution, expectations and performance, feedback, and assessment. These were domains identified in relation to observations made during the simulation assessments. Some domains and dimensions are closely linked or have some overlap, while others were independent.

This subsection may appear to the reader to be a predominantly negative reflection on the simulation scenarios. While observations made included some positive aspects of
the simulation assessment execution, the researcher’s focus and intention was to identify areas for improvement.

**Design**

The first main domain was *design*. This domain relates to observations made about aspects of the simulation scenarios that were inherent to the chosen format of the simulations and the simulation scenario instructions. These observations about factors at the design level are not affected by the staff (assessors, actors) or participants specific to this research project, and would be expected to be present if the same simulation assessments were run at another centre, or with different participants.

The first dimension identified within the design domain was *difficulty*. The suggestion that Simulation Scenario 2 was potentially more difficult than Simulation Scenario 1 was initially raised in the pilot phase of development. This difference in relative difficulty was considered to relate mostly to the portrayal of APO through history, and changes were made prior to the data collection period in an attempt to rectify this problem. However, it was apparent again in the data collection phase that Scenario 2 was more difficult, and it became apparent that this difficulty was inherent to the design of the scenario. The relatively higher difficulty of Scenario 2 was confirmed statistically in *Chapters 7 and 8*.

Within the first few simulation assessments, it became apparent that participants were struggling with the complexity of Scenario 2. This was first noticeable in the approach that participants took to the scenario. Participants who completed Scenario 1 apparently identified AMI as the most likely diagnosis early on (usually within the first one to two
minutes), realised the urgency of the situation, and took steps to confirm the diagnosis. Conversely, participants who were allocated to Scenario 2 were slower to recognise the diagnosis of APO from the initial primary complaint of dyspnoea. While participants did generally come to the diagnosis of APO, many failed to interpret the more complex underlying processes, that is, they did not necessarily understand that the underlying cause of APO in this scenario was AF, that the APO was exacerbated by the intravenous fluids that were given, and that the AF was triggered by infection (cellulitis).

It was also noticeable that some participants became lost in the amount of information available to them for Scenario 2. They seemed to struggle to reconcile the information in the patient notes (Appendix 5: Simulation Scenarios) with the current primary complaint of dyspnoea. In the debrief, participants fed back that they found it difficult to figure out what was happening in Scenario 2, whereas those who completed Scenario 1 reported that they found the scenario straightforward. Once the relationship between the different issues in Scenario 2 (APO, AF, excessive intravenous fluid, cellulitis) was explained to participants, they all understood the relationship and underlying pathophysiology.

Participants’ good understanding of the explanations given for Scenario 2 indicates that it was not a case of the content or level of knowledge being too difficult, but rather that the clinical context added complexity to this simulation. Arguably, though, junior doctors need to be able to manage complex clinical presentations in practice, and so should be assessed on their ability to do so. It is important, though, that the level of complexity is matched to the participant’s level of experience. This scenario is well matched to the level of management expected of intern and HMO2 level doctors (according to the Australian Curriculum Framework (ACF) for Junior Doctors), which
is the level for which it was designed. However, it would be very challenging for MD2 students.

In regards to the approach to the design of Scenario 2, it had been decided from the start of the simulation design process that Scenario 2 would depict APO. The details of the triggering/related conditions were added later to create a realistic story, in that APO generally requires a trigger, and worsening of AF is often also triggered by an acute illness.\(^{123}\) However, this also meant that, to fully understand the scenario, participants had to determine a sequence of events, rather than a single diagnosis. It is important to note, though, that participants could actually achieve full marks without making these connections. There were only four checklist items that in any way related to the underlying causes, though they did not necessarily require the participant to understand the relationship between the underlying causes and the current situation. Two items required the participant to obtain a history of symptoms associated with dyspnoea (palpitations, lightheadedness- these were clues to AF), one was interpretation of an ECG (which showed AF) and one required participants to cease the IV fluid (given the patient had APO). Essentially then, determining the correct diagnosis/diagnoses was not technically necessary in order for participants to receive full marks on the checklist (though it is unlikely that they would achieve full marks on the GAS). This is further discussed in Chapter 24.

Perhaps the decision to have the scenario set on the ward, as opposed to in the emergency department (ED), also made the situation more complex, as there was more information (admission notes, charts, etc.) for the participants to sort through, compared to if the patient was in ED and not yet admitted.
The other dimensions identified in relation to the design domain were consistency and fidelity. The simulation designs (especially the instructions to actors and assessors) ensured that there was very good consistency between the experience of different participants on each scenario. The only inconsistency that was expressly allowed and accounted for in the design, was that the patient could be either male or female, so that actors of both genders could be used. While one could argue that a male is the more typical AMI patient, there are no other implications of this minor design inconsistency. Detailed instructions for how patients and nurses should respond in a range of situations meant that there was no need for actors to improvise, such that there was very good consistency in the history and responses (including symptomatic response to treatment). Similarly, the examination findings to be given were included in the Simulation Scenario document, such that assessors could simply read these out. The initial vital signs and changes to vital signs shown on the bedside monitor were also pre-determined and included in the Simulation Scenario document, so there was no discretionary element to how these were to change.

The simulations were designed to be medium fidelity, using a hybrid set-up, based on the resources available. Participants, based on informal feedback received by the researcher, looked upon the use of medium fidelity simulation, particularly use of actors as patients and the presence of a dynamic bedside monitor, favourably. Participants commented that it was useful to have simulated patients, as they were able to actually take a history and perform an examination. They were also impressed with the realistic presence of a bedside monitor, rather than having to request vital signs from the examiners. It was anticipated that participants might look unfavourably on the inability to elicit examination findings from the actual patient (i.e. patients did not have clinical signs, examiners gave examination findings), however participants did not offer any
positive or negative comments on this element of the simulation design. It would be particularly useful to obtain more formal and in-depth feedback from participants regarding their experiences of the simulation assessments, as discussed in Chapters 25 and 26.

Execution

The second main domain identified was execution. This domain relates to the aspects of the simulation assessment that depended on the way in which the simulation scenarios were presented/performed to the patients, excluding aspects inherent to the simulation design. The first three dimensions relating to the execution domain are all interrelated: reproducibility/consistency, fidelity and actors.

The most commonly expressed negative feedback from participants was that it was confusing to have elderly patients portrayed by much younger actors. The scenarios depicted a 62 year old patient and a 76 year old patient. The ages of the patients were determined prior to finding actors to play the roles, and were chosen to make the scenarios reflective of common clinical situations. Unfortunately the volunteers who played the part of patients were aged 21-26 years old. For Scenario 1, the age of the patient was clearly stated in the scenario stem presented to participants prior to entering the simulation room. The age was also included on the patient record that was given to the participants. For Scenario 2, the age of the patient was not included in the scenario stem, but was included in the patient records. Nevertheless, some participants said that they had to consciously remind themselves that the patient was much older than was apparent, and in some cases they reported that they had difficulty considering diagnoses that would be very unlikely in young, apparently healthy patients. It would have been
ideal to have actors closer in age to the patients they were portraying, but as this was not possible, it may have been reasonable to make the actors appear older (at least to serve as a visual reminder to participants) by, for example, wearing a grey-haired wig.

In regards to the fidelity/realism of the simulation scenarios, apart from actors not appearing to be the age they were portraying, there was only one other potentially significant issue, which was difficulty in accurately portraying the increased work of breathing that would be expected with APO. While actors attempted to portray this, it is difficult to act short of breath and hyperventilate, especially for the entire ten-minute duration of the simulation. This may explain why some participants felt that Simulation Scenario 2 did not depict a very urgent situation, and therefore they did not seek senior assistance. In real clinical practice, apart from using objective vital signs as a marker for severity of illness, clinicians make a “rapid visual assessment to categorise the patient’s condition in terms of severity”.(123)[p4] If the patient looks only mildly unwell, clinicians are unlikely to feel an urgency to seek senior assistance.

Some moulage was used in Scenario 2 (make up applied to depict erythema) but further moulage was not used due to the complexity of changing between scenarios (since the two scenarios were used in an alternating pattern). Perhaps if the resources were available to run each scenario in a different room, with a different set of actors, then it would be possible to apply further techniques to improve realism, since there would be less time needed to re-set between participants. Even without these additional resources, it would be possible to use simple techniques (eg. wig/simulated grey hair) to assist in properly depicting the patient’s age.
Since the actors used were volunteers and not professional actors, there were also some issues with actors (both patients and nurses) inconsistently and inadvertently offering too much information. For example, in some cases, when the patient was asked about whether he had a past history of heart disease, he/she offered that he/she had a family history of heart disease, such that the participants elicited this information without specifically asking for it (and therefore scored a mark for this checklist item). Another example is that in some cases, when the participants noticed that there was IV fluid running (in Scenario 2), some nurses asked “do you want me to keep it running?”, thereby prompting participants to think about ceasing the fluids (again, a checklist item). There was also one occasion where the actor gave an incorrect response which potentially affected the assessment: the actor in this case reported that his pain had resolved after initial management, such that the participant would have no reason to give further GTN or morphine, therefore preventing them from scoring a mark for this checklist item. Overall, however, the actors performed well, and were able to portray the information provided in the Simulation Scenarios document.

Ideally, the simulations would use professional actors (preferably of an appropriate age) to play the part of the patients, however, this was not possible in this situation due to cost. It is unclear whether it would be better to use actors to play the part of nurses, as there is a clear benefit to using real nurses, who have the medical knowledge to respond to the doctors’ requests appropriately.

There were also some inconsistencies in the experiences of different participants. While the simulation scenarios should have had good reproducibility, given the detail included in the Simulation Scenarios document, in practice this was not the case, most likely due
to different actors being used for different sessions. For example, in Simulation Scenario 2, many participants asked the nurse if the patient was febrile. Answers given by the nurses varied, with some answering yes, some answering no, some giving an actual temperature value, though some nurses gave the first recorded temperature on the chart and others gave the most recent temperature (only two temperatures were recorded on the chart). This issue could easily be avoided by pre-specifying the way in which nurses are to respond to requests for each vital sign parameter.

There was also one case where there was an error in set up for Simulation Scenario 2, such that the patient’s bed was already sat up at the start of the scenario (whereas the patient should have been lying flat, since one of the checklist items was to sit the patient up as part of APO management). In this case the student researcher attempted to subtly lower the head of the bed, but this was noticed by the participant, and this prompted the participant to sit the patient up.

The final dimension relating to the execution domain is unexpected outcome. This dimension is actually in reference to the way in which a single participant’s simulation assessment ran. An individual completing Simulation Scenario 2 began his assessment by taking a history and performing an examination as expected for a patient presenting with dyspnoea. However, after requesting an ECG, the participant misinterpreted the ECG as representing a STEMI (AMI) and initially proceeded to management on the basis of this incorrect diagnosis. The issue that arose was that the actors involved in the simulation were not prepared for how to react to management based on incorrect diagnosis. The Simulation Scenarios document, with instructions for actors, included instructions on what should occur if management steps (including all management)
were omitted, but there was no provision for what should occur if incorrect management was instituted, as this situation had not been predicted. Fortunately, the participant involved in this case did not prescribe any management that would cause significant harm to the patient and, eventually, realised that the patient was in AF and fluid overloaded (it was not clear if he recognised APO specifically), and instituted some aspects of appropriate management. He did not, however, realise that his diagnosis of STEMI was incorrect until this was discussed in debrief.

The actors playing the part of patient and nurse dealt with the situation well, given that it was so unexpected. The nurse did not question the participant’s orders and executed them as instructed by the participant (though she realised that the diagnosis and management was incorrect, nurses had been instructed not to question orders, prompt or make suggestions). The actor required some prompting from the student researcher, as he was not sure what effect(s), if any, the instituted management should have on patient progress (who attempted to prompt using gestures not seen by the participant). Vital signs would not have been affected by the participant giving aspirin, and therefore were adjusted as per the protocol included in the Simulation Scenarios document for any steps in management that were executed (the participant gave GTN and metoprolol, both expected steps in APO management).

While this unexpected outcome of the situation was handled well by the staff involved in the simulation, this raises the question of what should be done in this situation. The main point taken away from this unexpected outcome was that there was not provision in the instructions for what should be done in such a situation. It might be appropriate, in future, to have a specific contingency plan in the Simulation Scenarios document,
outlining how to proceed if a candidate proceeds with incorrect management. It may also be useful to apply a specific marking penalty for incorrect diagnosis or incorrect management. This is discussed further in Chapter 24.

Expectations and Performance

The researcher’s field notes included many observations relating to participants’ expectations and performance. Both participants and assessors naturally had preconceived ideas of what to expect during the simulation assessments and, in some cases, there were differences between what was expected and what occurred.

In relation to participants’ expectations, the major finding was that some participants, based on previous experience, had difficulty differentiating what was expected of them in a simulation scenario, as compared to an OSCE. This was despite clear instructions to participants that read “This is a simulation, NOT an OSCE. You should perform as you would in real life: DO NOT explain your actions or address the assessors.” While most participants followed the instructions given, many participants attempted to explain to the examiners what they would do, rather than actually proceeding with the action(s). Some participants asked the assessors if they were allowed to request specific investigations or institute particular component(s) of management. In these cases, participants were simply advised that they could do whatever they would actually do, or whatever they wanted to do (since some participants were medical students and therefore had no real life experience in ordering investigations or instituting management). It was also made clear in the instructions provided to participants that they were taking on the role of a doctor, so while even medical students should have expected to be making decisions, it was clear that some of the more junior medical
students felt lost or uncomfortable with this, and this may explain why they were attempting to seek confirmation from the assessors.

Uniting the dimensions of participants’ expectations and performance, was the issue of participants’ expectations sometimes being quite different to assessors’ expectations. It was quite common that during debriefing, when assessors mentioned checklist items that participants had omitted, for participants to justify that they did not perceive that particular item to be of significance.

For example, for Scenario 1, participants were expected to ask the patient about the common cardiovascular risk factors. Many participants failed to address these items, and when this was mentioned to them during debriefing, they expressed that they didn’t think this was important in this clinical situation. While this part of the history was not required to reach a diagnosis in this case, cardiovascular risk factor assessment is generally an expected part of evaluation of likely acute coronary syndrome. Even in the emergency department setting, asking about risk factors assists greatly in risk stratification and determination of the likelihood of the diagnosis of myocardial infarction.(123)

Similarly, many participants argued, for both scenarios, that they didn’t think it was important to request a chest X-ray. However, a chest X-ray is part of standard first-line investigations for any patient presenting with chest pain or dyspnoea, and may differentiate between other diagnoses. A chest x-ray in the setting of chest pain could also suggest aortic dissection, which can be difficult to differentiate on history but would significantly change management.(123) While some checklist items may be
more pertinent than these in the clinical setting, completion of the above items are indicative of thorough patient assessment, and help to differentiate between participants whose performance is adequate and those whose performance is exceptional.

Perhaps the most surprising, and the most interesting finding of the qualitative study, was that many of the more senior participants (especially HMO2s) had a very different overall approach to the scenarios than the more junior participants. In general, HMO2 participants tended to approach the situation less sequentially. They ‘skipped’ some of the items that they perceived as less important at the start of the assessment checklists in favour of moving ahead to investigation and management. This difference may not have been distinguishable based on simulation assessment results, as many HMO2 participants found they had time left at the end of the simulation scenarios, and then went back to those items they had skipped. The fact that they went back to most of the items that they had initially not completed, if time allowed, suggests that their different approach was a conscious choice, rather than reflecting any deficit in their knowledge. A similar observation was made in a study by Lypson et al, in which “group differences [differences in performance between groups of differing levels of experience] were found for global ratings of performance but not for the individual aspects of technical skill”.(75)[p88] Similarly, research on OSCE assessment by Hodges et al. has shown that “On global scales, the experienced clinicians scored significantly better … but on checklists, the experienced clinicians scored significantly worse …”.(124)[p1129]

Medical students are initially taught to assess and manage patients in a very sequential manner. This creates structural frameworks with which novices can approach clinical situations. It also helps to minimise the risk of overlooking any feature of patients’ clinical situations or prematurely or incorrectly instituting management plans. The
additional steps of re-evaluating patients and prioritizing management act as safeguards: they are important so that if patients deteriorate, steps can be taken to get assistance and institute further management.

The two major frameworks for patient assessment and management that are taught to medical students are the *ABCDE (Primary Survey)* approach, and the *Hx/Ex/Ix/Dx/Rx* approach (which, for ease of reference, shall henceforth be referred to as the *sequential approach*).

The *ABCDE (Primary Survey)* approach is employed predominantly in emergency/trauma situations and guides simultaneous assessment and management in a fixed, sequential manner.(123, 125) Figure 22.1, shows the *Primary Survey* approach, based on various emergency medicine textbooks and guidelines.(123, 125, 126)

Surprisingly, especially since Scenario 1 was set in an emergency department, none of the participants in this research project were observed to use the primary survey approach. It is possible that, to some degree, they used this approach in their initial bedside evaluation of the patient, but did not verbalise or otherwise demonstrate that this is what they were doing. For example, it is possible that on entering the room they quickly assessed that the patient was alert and talking, and therefore deduced that they had a patent airway, sufficient breathing, were not in severe circulatory collapse (had adequate cerebral perfusion) and had normal conscious state– an overarching assessment of steps A to D in the primary survey approach. This rapid visual assessment may have occurred relatively subconsciously, and is appropriate in the context of real life clinical practice,(123) but has the potential to pose difficulties in the context of simulation, where actors will tend to appear “well”, and assessment is based on observed
The sequential approach is used across a range of fields of medicine and is more detailed/robust. This approach teaches full assessment prior to establishing a diagnosis and instituting management, and therefore is not designed for use in critically ill or unstable patients. (3, 103, 123) Figure 22.2, shows the sequential approach, based on general medical textbooks. (3, 103, 123) Most of the participants up to and including intern level, used this approach for the simulation scenarios.
Experienced clinicians, especially when managing more complex or seriously ill patients, find that in order to provide efficient emergency care, they must work in a less strictly sequential manner. (103, 123, 126) “Tasks that are typically carried out sequentially often have to be carried out in parallel with history taking, examination and initial resuscitation often occurring simultaneously.” (126)[p5] Studies have shown that expert clinicians are more focussed in their approach to history taking, and use different problem solving approaches to novices, which tend not to fit with, or be captured by, the standard checklists used for performance-based assessment (e.g. OSCEs). (124)

It was observed that, during the simulation assessments, the approach used by many of

---

**Figure 22.2: Sequential approach to assessment and management in medicine**
the HMO2 participants more closely approximated the expert clinician approaches described in various texts.\textsuperscript{(103, 123, 126)} That is, the HMO2 participants, and even some of the (apparently) more confident and competent intern participants, were observed to take a less stepwise, and rather more integrated approach, to patient assessment and management. They tried to ensure that time was not being wasted, and they tended to ask the nursing staff to complete various tasks (e.g., check vital signs, insert a cannula, take bloods, perform an ECG, administer medications, contact senior medical staff) while they proceeded to clinically assess the patient. This made their overall management of the patient more efficient, and some of the participants who took such an approach were able to complete the simulation assessment in a shorter time than was allowed (ten minutes), generally without compromising thoroughness.

Most of the participants who used an integrated approach performed well on the simulations, not necessarily completing the checklist items in the order in which they appeared (note again that no marks were allocated for the order in which tasks were completed), but completing all the critical items. They also tended to score well on the GAS component of the simulation assessment. This was not true for all participants who used such an approach, as some participants skipped key items (predominantly history items and some examination items), presumably because they considered them less important than other items (e.g., diagnosis and management/treatment items) and seemed to be focused on overall efficiency.

Sometimes participants came back to certain items after they realised that they had time left over at the end of the assessment, and so again, they scored well. However, it is unclear if they would have come back to these items in real clinical situations, where there is no allocated time limit and completion of initial management of one patient is
usually followed by moving on to the next patient waiting to be seen. Some participants chose not to come back to items that they had skipped (sometimes explaining during debrief that they had not considered those specific points to be of importance—see later discussion on assessment in this chapter); the assessment scores of these participants obviously reflected the fact that they had skipped key checklist items. This difference in approach taken by the most senior participants may at least partially explain why their performance on an objective assessment tool may not have always reflected their level of clinical expertise, a finding that has also been borne out by previous research.(75, 124)

The adoption of a more integrated approach, approximating that used by expert clinicians, by some of the HMO2 (and intern) participants, was not expected, given that at the time of the data collection, HMO2 participants had only had around 16 months of postgraduate clinical experience, of which only approximately 8 weeks would have been emergency department experience.

Figures 22.3, 22.4, 22.5 and 22.6, on the following pages, have been created based on observations of the approach to the simulation assessments taken by the more senior participants in this research project, in conjunction with the descriptions given by some textbooks regarding the approach of expert clinicians to clinical situations.(103, 123, 126) These figures attempt to depict how an experienced clinician might approach an emergency situation.

Figure 22.3 represents the way in which experienced clinicians are able to form an overall gestalt impression by considering all aspects of assessment together, rather than sequentially.(126) In this approach, multiple aspects of history, examination and
investigation are collected and combined. Together, these aspects of patient assessment are filtered for relevance, likelihood, risk and interrelationship, to form an overall impression of the patient’s situation — that is, their most likely diagnosis and the severity of their condition. Sometimes this overall impression will also include dangerous differential diagnoses that require further assessment (especially investigation) for exclusion. From this overall impression, the expert clinician is then able to decide on the best course of management for the individual patient.

Figure 22.4 demonstrates parallel patient assessment and management (tasks performed simultaneously), which is the approach generally used by experienced clinicians in emergency situations. (126) Initial assessment may reveal a working diagnosis or a list of possible diagnoses, for which the patient receives some form of treatment (initially resuscitation if required, or initial treatment), and then the patient is reassessed for progress and response to treatment. Re-assessment may lead to re-consideration of the diagnosis, or addition to/alteration of the treatment plan. Continuous re-assessment, re-consideration of diagnosis, and changes to management occur with ongoing evaluation of the patient’s status.
Figure 22.3: Expert approach: Gestalt impression

Overall impression

Figure 22.4: Expert approach: parallel patient assessment and management

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Diagnosis</th>
<th>Management</th>
<th>Re-assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Overall assessment</td>
<td>- Working diagnosis</td>
<td>- Resuscitation</td>
<td>- Deterioration/</td>
</tr>
<tr>
<td>- History</td>
<td>- Differential diagnoses</td>
<td>- Initial treatment</td>
<td>stability/improvement</td>
</tr>
<tr>
<td>- Examination</td>
<td></td>
<td>- Ongoing management</td>
<td>- Effect of management</td>
</tr>
<tr>
<td>- Investigation</td>
<td></td>
<td>plan</td>
<td></td>
</tr>
</tbody>
</table>
Figure 22.5 represents a more stepwise, but flexible, approach to assessment and management, which in some ways is quite similar to that depicted in Figure 22.4. The cyclical shapes in the figure depict the fluid order of the steps, which the clinician may return to if required. The first two steps, of initial impression and initial assessment, shown in yellow in the diagram, will generally not need to be revisited as much as the latter steps, which the clinician may return to several times. In particular, the step of ongoing re-assessment is shown with no clear ending to the circle, representing the fact that re-assessment should be continuous, and may lead to re-consideration of the working diagnosis or management plan at any point.

Figure 22.5: Diagrammatic representation of an expert clinician approach to emergency patient assessment and management
Figure 22.6 depicts how an experienced clinician may handle, or direct, several “arms” of patient assessment and management simultaneously. In this figure, the central component of the approach is ongoing assessment and management. In some cases, the senior clinician may choose to delegate the “arms” of patient assessment and management to other clinicians, while he/she acts as team leader, overseeing the ongoing management plan and patient assessment. It is important to note that in this figure, as in Figure 22.3, there is no stepwise approach. All “arms” of the model lead back to the central component of ongoing management assessment, which is informed by, and directs, each of the other components. For example, initial overall assessment (centre) may direct a focussed examination, which may then dictate the overall impression (centre) that the patient requires resuscitation.

![Diagram](image)

**Figure 22.6: Diagrammatic representation of an expert clinician approach to patient assessment and management in an emergency situation**
Comparison of Figures 22.1 and 22.2, to Figures 22.3, 22.4, 22.5 and 22.6, exemplifies the fact that the management approaches that are expected of junior medical staff, and which form the basis for assessment in clinical observation and simulation, are quite different to the optimal emergency management processes that are employed by experienced clinicians. Therefore, some deviation from the processes expected of junior clinicians might actually represent maturation of the clinician and development toward a more expert approach. Similarly, in some cases, lower marks on the checklist component of the simulation assessment may not necessarily reflect poorer clinical management, but rather an appropriate deviation from the approach to assessment taught in medical school.

Feedback

The feedback domain relates to the informal feedback provided by participants and assessors, rather than the feedback provided in the debriefing component of the assessment (which is included in the next domain, assessment). Please note that the lack of a formal tool for collecting participant and assessor feedback has been acknowledged as a limitation of the research, and this is discussed in detail in Chapters 25 and 26.

Most of the feedback received from participants in relation to the research project, and specifically the simulation component, was very positive. Participants appreciated the opportunity to partake in a simulation individually, and the individualised feedback and teaching they received.
Examples of some of the positive feedback received from participants include:

- Perhaps the most encouraging feedback that was received from a participant actually came months after the simulation sessions were conducted: One of the participants, an intern, reported that she had encountered a similar clinical situation to the simulation scenario and was required to initially manage the patient herself (i.e. there was no more senior doctor present). She said that, in assessing and managing the patient, she thought back to the simulation scenario and the feedback and teaching she had received, and this had helped her to manage the situation with greater confidence.

- Several students and JMOs mentioned that this was the only opportunity that they had ever had for individual participation and feedback in simulation scenarios, with previous experience always involving group management of scenarios. They felt that undertaking a simulation scenario individually (i.e. as the only initial decision-maker) was a useful experience in that it reflected true practice, especially on ward cover jobs.

- Participants reported that the individual feedback and teaching they received as part of the simulation assessment was particularly valuable.

The negative feedback received from participants mostly related to issues with fidelity of the simulation scenarios, especially the complaint that the use of actors much younger than the patients being portrayed was confusing. They also reported that the simulation felt ‘artificial’, and therefore it was difficult for them to perform as they would in a real clinical situation. Unfortunately, this reflects the fact that simulations are artificial reconstructions of real clinical situations and therefore, examinee
performance in simulated scenarios may not necessarily reflect how they would perform in clinical practice.\(^{(21)}\)

Examples of some of the negative feedback received from participants include:

- Some participants felt that the timeframe of the simulation scenario (ten minutes) was too short to be expected to re-evaluate symptoms, especially for Simulation Scenario 2.

- Some participants reported that they forgot that the patients were actually supposed to be much older than the actors portraying them. This confused them as they felt that the pathologies suggested by the history, examination, and investigation findings were unlikely in young patients.

- A few participants reported that they found it difficult to ascertain/evaluate the urgency of the clinical situations, given that the patients (actors) appeared very well. As such, despite findings suggesting that urgent management was required, their overall impression was that (especially in Scenario 2), they could take their time to further assess and manage the patient.

Assessors offered less feedback than participants (perhaps because there were fewer assessors than participants) but most of the feedback that was received actually related to use of the assessment tool. Assessors generally reported that the assessment tool was easy to use due to its format and predominantly checklist marking. However, assessors also reported that it was sometimes difficult to make a dichotomous choice as to whether to allocate a mark for a checklist item or not, as they sometimes felt that participants had partly fulfilled a criterion, or sometimes performed an equivalent action.
that was not listed on the assessment tool. For example, for Scenario 2, some participants used amiodarone or digoxin to manage the atrial fibrillation, which are also reasonable approaches,(123) though initiation of management with these drugs is generally considered to be beyond the scope of an intern or HMO2 level doctor. Metoprolol was included on the checklist in preference to amiodarone or digoxin because the patient in the scenario was already on metoprolol, therefore giving more metoprolol (or giving the next dose early) would be within a junior doctor’s scope of practice.

**Assessment**

The final domain that was identified was *assessment*, with dimensions of *formative assessment* and *assessment tool*. As discussed previously, participants reported that they found the feedback and teaching provided during debriefing very useful. This is consistent with earlier findings of this research project, relating to the high formative validity of the simulation assessments. Qualitative analysis also uncovered that, in assessing the examinees, assessors were able to obtain very good insight into the strengths and weaknesses of each participant and the overall strengths and weaknesses of the group of participants. This again confirms the finding that the simulation assessments had high formative validity, since, as reiterated throughout this thesis, formative validity also relates to the ability of an assessment to guide further teaching and learning.(16, 51)
The researcher’s field notes included many observations about the simulation assessment tools, all of which related to issues and potential improvements that could be made. Findings that related to specific assessment items included:

- **Scenario 1, Item A3a**, which related to obtaining history of cardiovascular risk factors: This item was allocated two marks but perhaps this overestimates the relative importance of this item, and one mark would be more appropriate.

- **Scenario 1, Item A5e** and **Scenario 2, Item A3e**, which related to examination of the legs: These items did differentiate between examining for peripheral oedema and examining calves for signs of DVT. Some participants only examined for one of these findings, while others examined for both, yet received the same mark. It might be appropriate to allocate separate marks for each form of leg examination.

- **Scenario 1, Item A6c** and **Scenario 2, Item A4d**, which related to requesting blood tests: Minimum blood tests required were included in the criterion, but some participants requested only one/some of the investigations listed in the criterion (e.g. some participants requested only troponin, but not FBE & UeC for Scenario 1). There may need to be a provision for assessors to allocate half a mark if only some of the blood investigations are requested. Alternatively, a mark could be allocated only for one key blood investigation.

- **Scenario 1, Item B1c** and **Scenario 2, Item B1b**, which related to giving oxygen as part of management. There may need to be a threshold defined for a minimum amount of oxygen considered therapeutic. With the current wording of the criterion, a mark would be allocated if the participant gave only very low flow oxygen via nasal prongs to treat significant acute dyspnoea in APO.
• Scenario 1, Item B1d, which related to giving GTN or morphine to treat ongoing pain. Many participants initially gave GTN and morphine, before evaluating for ongoing pain. It is unclear whether this would fulfill the current criterion. It may be more suitable to allocate one mark for giving GTN and one mark for giving aspirin, since there is already an item allocated for re-evaluating pain severity (item B2a).

There were also some more general issues identified in relation to the assessment tools. Firstly, the physical format of the assessment sheets was problematic in that for all checklist items, assessors were required to circle either Y (yes, i.e. criterion completed), or N (no). However, assessors tended to circle (or tick) Y, but not N. Presumably assessors did this so that they could leave any criteria that weren’t completed until the end of the simulation, in case the participant returned to that item later. In general, though, the assessors did not go back to mark N for these items, so it had to be assumed that the lack of a Y implied that the item was not completed. Given this finding, it may be more appropriate to just have a tick box for each criterion, where the assessor can tick if completed, and leave blank if not completed. In this way, the marking is not open to interpretation.

The other possible issue identified in relation to the assessment tool was the descriptors used for the global assessment scores (GAS). Some assessors complained that these were too rigid, however it is important to have clearly defined criterion-based definitions of the different levels for the GAS, so that subjectivity is reduced. It is also important to ensure that the definitions given for each GAS level (score 0-3) do not overlap, such that assessors are able to confidently choose the score that best matches the candidate’s performance, rather than feeling that the candidate’s performance could
fall in either of two (or more) categories.

This chapter has addressed Research Question 4- “What are the strengths, weaknesses, limitations and issues related to conducting simulation assessment research?” The following section, Section V: Summary and Conclusions will provide an overview of all the pertinent findings of this thesis, and the implications of these findings.
SECTION V: SUMMARY AND CONCLUSION

CHAPTER 23: SUMMARY OF THE RESEARCH FINDINGS

A. Participant Characteristics

Thirty-eight participants were recruited for this research project: 17 students and 21 JMOs. The mean age of participants was 25.8 years, and 63.2% of participants were female. Most participants had completed a graduate-entry medical degree, and the majority had completed their medical training through The University of Melbourne. Thirty-seven of the participants completed one of two simulation assessments. There were no statistically significant differences in the demographics of the groups allocated to the two simulation scenarios.

B. MCQ and Simulation Assessment Results; Correlation Between the Two Assessment Types

These findings address Research Question 2- “What is the relationship between performance on simulation-based and MCQ assessments of cardiac emergency management?”

The mean score for the MCQ test was 22.95 (SD 3.44). The mean scores for the Scenarios 1 and 2 were 22.89 (SD 2.94) and 16.06 (SD 2.67), respectively, demonstrating that Scenario 2 was more difficult than Scenario 1. There was a strong and statistically significant correlation between the checklist and GAS components of Simulation Scenario 2 ($r = 0.71, p = 0.001$), however the correlation between the two components of Simulation Scenario 1 was low and not statistically significant ($r = 0.41, p = 0.082$). This is likely accounted for by the subjectivity involved in determining the GAS, though the reason for the greater correlation between the checklist and GAS...
components for Scenario 2 is unclear; perhaps since participants overall performed better on Scenario 1, there may have been greater difficulty in differentiating their performance using the GAS. There was a moderate correlation between MCQ and simulation assessment results ($r = 0.52$ for Scenario 1, $r = 0.57$ for Scenario 2), suggesting that the two different assessment methods were assessing similar constructs.

C. Content Validity

These findings address Research Question 1E- “What is the content validity of each method of assessment?” and partially address Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?”

Qualitative analysis demonstrated that both the MCQ test and simulation assessments had adequate content validity, but that the simulation assessments more evenly addressed the different learning outcome categories. This was attributed to the practical nature of the simulation assessment tool, which therefore allows assessment of all facets of clinical management.(22) The MCQ test assessed the content of both simulation assessments combined (AMI and APO management), in the same time that would be taken to run both simulation assessments (30 minutes). However, as discussed in detail, the time taken to assess a group of examinees would be much shorter using the MCQ test, since multiple examinees can be assessed at the same time with this method, whereas simulations are undertaken individually.
D. Construct Validity

These findings address Research Question 1A- “What is the construct validity of each method of assessment?” and partially address Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?” It was difficult to interpret the various findings in relation to construct validity, since different measures of construct validity yielded different findings. Internal consistency, as measured by coefficient alpha, was unacceptably low for each of the three assessment tools, but higher for the MCQ test ($\alpha = 0.64$) than for Scenario 1 ($\alpha = 0.52$) or Scenario 2 ($\alpha = -0.03$). Construct validity, as measured by the contrasted-groups approach, was proven for the MCQ test and Scenario 2, but not for Scenario 1. None of the assessments were able to show consistent improvement across the range of classification. Overall, based on the findings of all measures, the MCQ test appeared to have the best construct validity.

E. Formative Validity

These findings address Research Question 1B- “What is the formative validity of each method of assessment?” and partially address Research Question 3- “How do the psychometric properties of the MCQ test and the simulation assessments compare?”

Quantitative methods demonstrated that all three assessment tools had a range of mean item values across the different learning outcome categories, suggesting that they provide information on performance of all candidates, including details of areas of strength and weakness. Qualitative evaluation demonstrated that the MCQ test had adequate formative validity, limited by the inherent properties of this assessment method. The simulation scenarios had excellent formative validity, providing both
assessors and examinees with a large amount of detailed and valuable information to guide further teaching and learning. In addition, the simulation assessments were able to provide more information about the candidates’ performance in the same amount of (individual) testing time, because they included a larger number of assessment items.

F. Interobserver Reliability

These findings address Research Question 1D- “What is the inter-observer reliability of cardiac emergency simulation as an assessment tool?”

Interobserver reliability was very high for Simulation Scenario 1 and fair for Scenario 2. Individual items with higher interobserver variability were identified; if these items were improved, the overall interobserver reliability would be expected to improve. Another approach to improve interobserver reliability would be to provide training to assessors.(116)

G. Research Process

These findings address Research Question 4- “What are the strengths, weaknesses, limitations and issues related to conducting simulation assessment research?”

Qualitative analysis was undertaken on the topic of participant recruitment and logistics in order to identify factors influencing these aspects of the research project. The major domains identified were time, resources, interest and communication. Dimensions within these domains and areas of overlap were also identified. These findings helped to identify barriers that led to lower than target recruitment, such that these could be addressed for future research.
Qualitative analysis was also undertaken to evaluate observations made in relation to simulation assessment execution. This uncovered various findings about the simulation assessments that were not identified during the process of evaluation of psychometric properties, but may be useful to consider as part of a detailed evaluation of the simulation assessment tools. Findings related to five major domains: design, execution, expectations and performance, feedback, and assessment.
CHAPTER 24: RECOMMENDATIONS

1. **Simulation should be used as a formative assessment method to evaluate the ability of medical students to manage cardiac emergencies prior to graduation.**

   Junior doctors are expected to be able to manage cardiac emergencies with confidence and efficiency. (4-7) As such, it is incumbent on medical schools to ensure that their graduating students are able to provide good quality medical care in cardiac emergency situations. (5, 8) This research project has demonstrated that the simulation assessments that were evaluated have excellent formative validity, and are able to provide very useful information to both educators and learners about overall performance, as well as areas of strength and weakness. While written examinations also have advantages, it seems logical that at some point, graduating medical students should be assessed on their ability to manage clinical situations with a practical assessment method that approximates clinical practice. (59) This research project has demonstrated that simulation-based assessments can be used as a reliable and accurate assessment method for this purpose.

2. **Due to the differences in the psychometric properties and optimal uses for different assessment methods, medical schools consider using a range of different assessment approaches.**

   It is generally accepted that knowledge is best assessed by written assessment methods, such as MCQ tests and short answer tests, (11) while practical skills are best assessed using a practical assessment method. (22) This research project has expanded on these overarching principles by demonstrating, through comparison of an MCQ test and simulation assessments designed to assess the same content, that each assessment method, and each individual assessment tool, has strengths and weaknesses across different psychometric domains.
In choosing an assessment method, it is important to consider which psychometric properties are important for the specific situation. For example, if the assessment were to be used for formative purposes, then the assessment method with greatest formative validity would likely be preferred. For assessments undertaken very early in medical school training, before there has been much clinical exposure, the aim may be to assess knowledge rather than practical application of knowledge, in which case, an MCQ test would be preferred. (11) The findings of this research, in regards to the psychometric properties of the different assessment types, are not necessarily generalisable to all assessment tools of the same type (MCQ or simulation), but demonstrate the concept that different assessment approaches have different psychometric properties, and therefore should be considered to have different educational applications. It follows, then, that to ensure robust assessment during a course of medical education, it would be advisable to use a range of different assessment approaches. Of course, feasibility of the assessment tools in the specific educational context would also need to be considered. (11, 16)

3. Educators should evaluate the assessment tools they use, in terms of psychometric properties, both prior to their use for assessment purposes, and ongoingly. They should use the findings of these evaluations to choose the most appropriate assessment tools for specific circumstances, and to make alterations to the assessment tools, to improve their reliability and validity.

Ideally, assessment tools should be valid, reliable and feasible. (11, 50) However, it is not necessarily possible for an assessment tool to fulfil all of these criteria. Therefore, compromise in some psychometric properties may be necessary. (11) It is then important to be aware of the limitations of any assessment tools being used, in terms of the various types of reliability and validity. (50) This research project has demonstrated
that evaluation of assessment tools may yield complex, mixed findings.

In this research project, the two simulation assessments yielded different results across some psychometric properties, despite being designed very similarly, using assessment sheets based on the same structure (modified for the clinical scenario) and being applied to participants drawn from the same cohorts. This adds further weight to the recommendation that educators should evaluate the specific assessment tools (not just type of assessment) they use in the context of their specific educational setting and population, to ascertain whether they are psychometrically robust, and to choose the most appropriate assessment tool for a given circumstance. (50) While there may be previously reported reliability and validity information available for some assessment tools, these findings would not necessarily hold constant in a different context. (50)

As discussed in Chapter 19, a small number of items likely contributed to the low internal consistency of the MCQ test used in this research, as measured by the coefficient alpha value ($\alpha = 0.64$). Analysis of individual items did not reveal any items that, if removed, would significantly alter the alpha coefficient for the entire assessment. However, this value would be expected to improve with the addition of further items, so long as the additional items correlate well with the existing items. (50, 52, 88, 89)

This demonstrates the principle that MCQ tests (and other assessment tools) should be piloted and evaluated for various psychometric properties, such as internal consistency, prior to use in assessment. (79) This may allow educators to make simple changes, such as the addition of further assessment items to improve internal consistency, or deletion of items with poor content validity, in order to enhance the overall utility of the
4. The fidelity/authenticity of simulations should be considered in the execution and assessment of simulation scenarios.

Inauthentic representations of clinical scenarios detract significantly from the utility of simulations, in that the purpose of simulation is to replicate real life clinical practice. (10, 34, 47, 127) Lack of authenticity may, at best, result in poor replication of what the examinee would do in clinical practice (e.g. low fidelity procedural mannequins); at worst it may distract or confuse the examinee to a point where they either form an incorrect working diagnosis, or become disengaged. (10, 28, 127)

In this research project, resource constraints meant that untrained actors who did not fit the description of the patient (were much younger) were used. This significantly affected the fidelity of the simulations and, in some cases, as discussed, possibly affected examinees’ decision-making. Ideally, simulations should be performed with the highest possible level of fidelity. Where there are obvious issues with fidelity, these should be clarified with examinees (e.g. examinees in this research project were clearly told the patient’s age, and the patient’s date of birth was also written on the patient files). These issues should also be taken into consideration when assessing performance. (21) For example, where there are equipment/technology failure issues, examinees should not be penalised for consequential errors.
5. Simulation assessment checklist items and global assessment score criteria should be clearly defined and unambiguous to reduce the subjectivity in assessment and improve interobserver reliability. Marking penalties may be appropriate for incorrect diagnosis or management. Performance-based assessments are generally scored using a checklist approach, with or without an additional global rating, in order to minimise potential subjectivity and therefore enhance interobserver reliability. However, the ability to score objectively depends on having clearly defined criteria for each checklist item and for the global assessment score.

It is important that the requirements to meet each criterion are clearly defined, so that achievement of the criterion is not open to interpretation by the assessors, and, importantly, also not able to be argued by the examinee. Checklist items should be marked dichotomously (achieved/not achieved) or according to a clearly defined scale (if marked non-dichotomously, e.g. on a three point scale). The scoring for the global assessment score must also be criterion-based, with each level of achievement clearly defined, avoiding vague or general descriptors, and with no overlap between the defined levels of the scale.

The scoring sheets used for the simulation assessments in this research did not require the participant to reach the correct diagnosis in order to achieve full marks on the checklist component, though it is unlikely that a participant who made an incorrect diagnosis would score full marks on the GAS component. This issue could be addressed by allocating specific marks (perhaps more heavily weighted than checklist marks) for correct diagnosis, though this would still allow participants to mismanage patients without penalty. In order to emphasise the importance of reaching a correct diagnosis
and managing patients appropriately, it may be appropriate to apply marking penalties for incorrect diagnosis and/or subsequent management (especially potentially dangerous management). This may require assessors (or the nurses in the scenario) to specifically ask the examinees for their working diagnosis.
CHAPTER 25: LIMITATIONS

There were a few limitations that impacted the ability of this research project to generate more definitive or generalisable findings.

1. Sample size

The small number of participants recruited overall, and the very small number of participants within the different classifications, means that there is greater likelihood of sampling error. That is, that the statistics obtained based on the results of the sample population, overall and for any cohort/classification, will not represent the population from which the sample is drawn. (50)

With very small numbers of participants of a given classification allocated to each of the two scenarios, and sometimes even no participants (only one of the eight MD2 participants was allocated to Scenario 1, none of the three MD4 participants were allocated to Scenario 2), sampling error was likely implicated in the inability to show consistent, or statistically significant, improvement across the range of classifications, when examining for construct validity. Sampling error due to small sample sizes may have also been implicated in the inability of individual test items to distinguish between examinees of different classifications, one of the quantitative measures used to evaluate formative validity.

Small sample size also meant that, in one situation in particular, the most appropriate statistical analysis method could not be used. Specifically, in examining construct validity using the contrasted-groups approach, since there was only one MD2 participant allocated to Scenario 1, the MD3 group instead was used to represent the ‘novice’ category. This test was then unable to prove construct validity for Scenario 1,
and it is unclear if this is because the assessment tool truly lacks construct validity, or because an inferior statistical approach was used. Similarly, not all assessors’ marks could be compared using all measures of interobserver reliability, due to the prohibitively small number of cases assessed by some of the assessors.

Statistical analysis was not performed on some of the data obtained from the participant questionnaire (pertaining to previous education, previous experience), as the subgroups were too small to obtain meaningful and interpretable results.(128, 129)

Sampling size, therefore provided some limitations in addressing Research Question 1A- “What is the construct validity of each method of assessment?”, and Research Question 1B- “What is the formative validity of each method of assessment?”

The reasons for poor recruitment were identified in the qualitative analysis of participant recruitment and logistics (See Chapter 16 and Chapter 22). Some of the most significant issues identified were related to:

- resources (physical, human, time)
- logistics (coordinating time and resources)
- competing priorities of participants

Ideally, recruitment could have been improved by extending the recruitment period, or offering participants the chance to participate during their protected teaching time (eg. intern teaching). Unfortunately this was not possible on this occasion, due to time and human resource constraints, and the fact that the simulation centre was in the process of being moved to another location at the time that the research was being undertaken (such that the end of the data collection period could not reasonably be extended).
Another potential method to mitigate the issue of poor recruitment in the future would be to have participants complete both simulation assessments, as discussed in the next chapter.

2. Non-random allocation

Non-random allocation is included as a potential limitation, although there is no evidence that use of alternation, rather than randomisation, had any negative impact on the study. Alternation (allocation of participants alternately to the two simulation scenarios) was used for this research project to simplify the process for actors and assessors. It was considered that it would be easier for set up to alternate between the two scenarios, such that there was no need to think about which scenario was next, and the actors and assessors could simply switch between the two.

Alternation, as a method of systematic allocation, should result in even allocation of participants to groups, as would occur with randomisation, however neither method is likely to produce a perfectly representative sample. Ideally, to produce the most representative samples, stratified sampling should be used. Stratified sample involves allocation, either random or systematic, of participants from within subgroups (in this case, classification would be the most important subgrouping). This should result in the allocated groups being representative of the larger group (the entire study population). In retrospect, this approach could have been used, despite the fact that recruitment and data collection occurred simultaneously, such that the number of participants from each classification was not known in advance. If participants from each classification, rather than all participants, were allocated by alternation (i.e. every other MD2 student was allocated to Scenario 2, etc.), then participants from each classification would have been evenly divided between the two scenarios.
Use of stratified sampling, instead of alternation, would have avoided some of the problems related to small sample size, and may have improved the ability of the research study to address Research Questions 1A and 1B.

3. Multiple assessors marking small numbers of cases, no assessor training

Apart from Assessor 1, who scored all 37 simulations, the other assessors changed frequently, due to availability, and therefore scored only a small number of cases. Only Assessors 3 and 5 scored enough simulations (fifteen and twelve, respectively) to properly assess interobserver variability, but even these numbers are small, considering that they were divided between assessments of Scenario 1 and Scenario 2.

Ideally assessors should be trained prior to assessing candidates because, as previously mentioned, interobserver reliability improves with training of assessors. (116) Unfortunately, formal in-depth training was not practicable for this research project, due to time constraints and, more importantly, a lack of example cases to use to practice marking. As an alternative for formal training, assessors were given the opportunity to familiarise themselves with the assessment marking sheets prior assessing, and a verbal explanation of the assessment and debrief process was also given by the student researcher.

A possible method to train at least some assessors using minimal resources would be to allow assessors to observe and practice marking a few cases prior to actually assessing candidates. The number of training cases required for assessor training has not been well established, and may depend on a number of factors, such as assessor experience, familiarity with similar assessment tools, familiarity with the simulation content, understanding of the place of the assessment within the curriculum and the expected
standard of trainees, and the form of assessor training (eg. online, face-to-face).

Regardless, it is well accepted that assessor training, of any description, will improve standardisation/consistency of marking.(116)

A possible solution to the problem of each assessor only marking a small number of cases would be to record the simulations, thereby allowing multiple assessors to review each simulation remotely (and allowing multiple assessors to assess each simulation/examinee). As discussed in Chapter 3, a decision was made initially not to record the simulations, however, it would be appropriate to reconsider this decision for future projects. Similarly, a recording of a pilot simulation run-through (or previous simulation assessment, if the same simulation is used again) could be used as an orientation and training tool for assessors, to demonstrate how the simulation will run and discuss how the assessment tool should be employed.

These issues in relation to assessors may have impacted the ability of the study to fully address Research Question 1D: “What is the inter-observer reliability of cardiac emergency simulation as an assessment tool?” The research question has still been addressed, but the low sample size clearly affects the generalisability of the findings.

4. Formative validity of the assessments was limited by participants not receiving their results

As discussed in Chapter 20, a decision was made not to provide participants with their assessment results. This was in the context of prioritising maintenance of the integrity of the assessment instruments for research purposes (i.e. to limit sharing of answers between participants, which would have affected results). Ideally, and in the setting of true assessment (e.g. as part of a course or for credentialing purposes), participants
should receive, at the very least, their overall score for an assessment. This provides some low level feedback, which can assist trainees in guiding their future study and development. In retrospect, this information should have been provided to the research participants, and this would have enhanced the formative validity of the assessment tools.

5. Lack of depth and breadth of qualitative data

All the data forms used for the qualitative component of the research were in some way created by the researcher: the field notes were based on the observations of the researcher or discussions between the researcher and other ‘players’ in the research (participants and assessors); the other data forms analysed (assessment tools, Multiple Choice Question Test Feedback document, debrief instructions for the simulation assessments) were designed and developed by the researcher.

Ideally, the qualitative analysis component of this research would involve use of some ‘external’ qualitative data, for example, data gathered directly from participants, assessors, or even others not involved in the research (educational experts, etc.). This is a limitation of the methodology used in this research. The use of a qualitative data collection tool for participants and assessors was not considered in the initial study design. The inclusion of a tertiary objective, examination of the research process through a qualitative approach, was a later addition to the study design in response to recognition of difficulties encountered with practical aspects of the research. By this time, it was not possible to add additional participant and assessor data collection tools, since it was not feasible to seek changes to ethics approval from each of the organisations involved. As such, it was decided that a pseudo-fieldwork approach, and qualitative analysis of the research materials (assessment tools, etc.) would be used.
Suggestions for additional qualitative data collection methods are discussed in Chapter 26.

When the limitation in the sources of qualitative data was recognised, steps were taken to attempt to control for the single source of data (the researcher) and potential for researcher bias. Firstly, the incorporation of informal feedback from both participants and assessors meant that the researcher’s field notes actually included data from other sources, and not just the researcher’s observations. Specific steps were also taken to control for potential researcher bias. For example, when coding the field notes, feedback from participants was coded specifically as positive and negative (as well as being coded/tagged as feedback), and particular effort was made to include both positive and negative participant views in the discussion surrounding the relevant domains and dimensions. Similarly, examples of specific feedback given by the participants included both positive and negative comments/observations. It is hoped that these methods of addressing the lack of external qualitative data at least partially addressed this limitation.
CHAPTER 26: FUTURE DIRECTIONS

This research has provided some useful information on assessment of competence in management of cardiac emergencies at a medical student/junior doctor level, an area with very little published research. However, further research is needed to establish the strengths and weaknesses of the different methods of assessing cardiac emergency management and, ultimately, determine the optimal assessment methods for this purpose.

Building on this research project, future projects could follow the same methodology that was used for assessing the psychometric properties, but address the limitations identified and potentially build on the findings by:

1. **Having participants complete both simulations (or multiple simulations)**

   Having participants complete both (or multiple) simulations would offer benefits both in terms of research data as well as more comprehensive assessment. In terms of enriching the research data, participation in both simulations would allow for direct correlation between performances on the simulations, as well as increasing the sample size for each simulation assessment. An additional benefit would be that participants would have more opportunities for learning by participating in more simulation scenarios.

   Unfortunately, in this case, once the issue of low recruitment had been identified, simulation sessions had already begun and it was not possible to amend the ethics applications at all institutions in time for the process to be changed for later participants. (All simulations were run within a one-month time frame.) Similarly, it was not possible to extend the time frame over which the simulations were run (the simulation
centre was being relocated at the time), so there was limited scope to address the issue of low recruitment.

In terms of enhancing assessment, if used in the context of formal assessment, participation in multiple simulations would provide a greater amount of information about the candidates’ competence. In this research, participants only participated in one simulation each, as the aim was to validate and compare assessment tools, rather than to comprehensively assess a participant’s progress or readiness to progress to the next stage of training (as would be the aim in medical school assessments). In order to draw conclusions about a candidate’s competence, it is necessary to assess performance on multiple simulated patient encounters. The literature suggests that ten stations (simulations), or three to four hours of testing time results in good reliability in OSCE assessment (including clinical simulation stations). (21, 25)

2. Adding a qualitative data collection tool for participants and assessors

Adding data collection tools to gather specific qualitative data from participants and assessors (e.g. surveys or focus groups) would add both depth and breadth to the qualitative component of this research. In particular, this would allow more formal feedback to be obtained from participants, especially in relation to the examinees’ perspective on the formative validity of each assessment method. It could add another level to the research, by seeking participants’ more general opinions on the different assessment methods, for example, their views on the relevance of each assessment method (essentially evaluating assessment acceptability). Similar information could be obtained from the assessors, which would add much greater breadth to the limited data obtained through a single researcher’s observations.
Qualitative data collection tools for participants and assessors could also be used to obtain further data in regards to the research process. For example, analysis of participant recruitment issues was based on observations and informal feedback, whereas the perspectives of the participants would be particularly useful here (though input from those who chose/were able to participate would obviously be biased). Similarly, while informal feedback obtained from participants and assessors regarding the simulation assessment execution was very valuable, further exploration of their perspectives through more formal data collections would allow for much broader and deeper analysis of the phenomenon of simulation assessment execution.

3. Adding a longitudinal component

The scope of this research could be expanded by addition of a longitudinal component; re-testing participants at a later date, using one or both assessment methods, could be used to determine whether the simulation experience(s), including debrief, had any educational value. This would also help to evaluate further the formative validity of the simulation assessments. In order to attribute any change to the simulation experience, there would need to be a control group that was not exposed to the simulation experience between the initial and repeat assessments. It could also be useful to obtain feedback from participants about whether they felt that the simulation experience had any impact on their clinical practice or level of confidence.

4. Assessing the impact of the order of assessments on performance

Completion of the MCQ test prior to participation in a simulation assessment may have impacted participants’ performance on the simulation, in particular by reminding participants of aspects of investigation or management that they may not have otherwise thought of. This does, however, reflect common end of semester assessment practices,
where students may sit their written exams prior to, and within close temporal proximity
to, their practical examinations (e.g. OSCE).

Ideally, in order to examine and account for the effect of the order of assessments, half
the participants would undertake the MCQ prior to the simulation, and half would
undertake assessments in the reverse order. Prior to commencing this study, it had been
decided that if participant numbers exceeded approximately eighty, it would be
reasonable to further divide the participants in order to have half complete the MCQ
first and half complete the simulation first; dividing a smaller number of participants
would result in groups too small to properly analyse. As recruitment did not reach this
number, it was not possible to examine the effect of the order of assessments on
performance.

5. Future related research

Avenues for further related research could include:

1. Undertaking similar research in different areas of emergency medicine- e.g.
   trauma, obstetrics/gynaecology/paediatrics

2. Comparing other assessment methods (e.g. traditional OSCE, workplace-based
   assessment, short answer written test)
CHAPTER 27: CONCLUSION

This thesis has presented the findings of a research project designed to validate simulation as a method for assessing competence of medical students and junior doctors in managing cardiac emergencies. Given that MCQ tests are commonly used for assessment in medical schools, the MCQ test was used as a comparison with which to compare the simulation assessments. Three different assessment tools were designed for the purpose of this research: an MCQ test assessing AMI and APO management, and two simulation assessments, one assessing AMI management and the other assessing APO management. These assessment tools were tested on a sample population of medical students and junior doctors at Northern Clinical School/Northern Health to evaluate and compare their psychometric properties.

Overall, the two simulation assessment tools designed for this research project were shown to have similar psychometric properties to the MCQ test. Each assessment method had strengths and weaknesses, but a notable difference was the superior formative validity offered by the simulation assessments. This, however, was a relatively subjective finding, which requires further confirmation with longitudinal testing, as suggested in the preceding chapter.

There was moderate correlation between results on the MCQ test and simulation assessments, therefore it would be reasonable to consider using simulation assessment in medical school assessment, in addition to the traditional MCQ assessment. In particular, simulation would allow assessment at a higher level of Miller’s pyramid, that is, performance. Simulation would be especially useful as a formative assessment, to allow examinee results to guide ongoing teaching and learning. However, simulation could also be used as part of summative assessment, so long as the essential debrief
MCQ tests are not only proven by decades of research across many fields, but they offer benefits such as being useful to assess lower levels of knowledge in isolation (as may be required in the earlier stages of medical school), less resource intensive (financial, human and time resources) and easier to mark. However, simulation-based assessment is the only assessment method (apart from observation of clinical practice) that provides the opportunity to directly test students’ application of knowledge and their ability to actually assess and manage patients. The realism of simulation also provides a unique opportunity for students to experience the pressure and anxiety of managing an emergency situation (though some habituation is expected), without compromising patient safety.

At this point, based on this research and previous research investigating the psychometric properties of simulation-based assessment, it is reasonable to recommend that medical schools consider using simulation-based assessment in conjunction with, but not instead of, more traditional assessment methods, such as MCQ tests. As previously stated, different assessment methods have different psychometric properties, and no single assessment method can provide adequate reliability and validity across all psychometric domains. Similarly, no assessment method can comprehensively assess performance across all knowledge/performance/skill domains. As such, it is appropriate to use a range of assessment methods, including simulation-based assessment, to assess the breadth, depth, and application of knowledge that we expect of our graduating medical students.
This research has demonstrated that simulation may be used to assess management of cardiac emergencies, a finding which is also supported by the research by Opar et al. (71) Other research has demonstrated the utility of simulation-based teaching and assessment in other aspects of medical (and nursing) education, including interprofessional practice, emergency medicine, military medicine, surgical skills and resuscitation. (8, 12, 34, 42, 71, 73, 74, 76, 130) Further research is needed prior to employment of this assessment approach more broadly in medicine.

It is important to also consider that the feasibility of introducing simulation-based assessment into medical education will depend on the resources available to the university and/or clinical school. As previously discussed, simulation is resource-intensive in terms of financial, physical and human resources, to the point where the resource requirement may be prohibitive in some cases, or simply mean that more traditional assessment methods are used in preference due to feasibility. (41) As such, it is important that there is ongoing research into the utility of this simulation as an assessment method, in order to justify its use.

This study was able to address all of the research questions that had been posed, though with some limitations, as discussed. Further research, on a larger scale, would assist in providing further validation for the use of simulation assessments for assessing cardiac emergency management.
REFERENCES


simulation training in emergency medicine designed in the light of malpractice cases. BMC medical education. 2014;14:155.


43. Newble D, Reed M. Developing and Running an Objective Structured Clinical Examination (OSCE). University of Sheffield; 2012.


47. van der Vleuten CPM, Newble DI. How can we test clinical reasoning? The Lancet. 1995;345:1032-4.


80. Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: Developing taxonomy, themes, and theory. Health Services Research. 2007;42(2):1758-72.
113. Jonas-Dwyer D, Siddiqui Z. Feedback in Higher Education: A Literature Review. Western Australia: University of Western Australia: Faculty of Medicine, Nursing and Health Sciences, 2006 August. Report No.


APPENDICES

Appendix 1: Participant Information Form and Participant Consent Form

Appendix 1: Page 1 of 7

Participant Information Sheet/Consent Form

This is based on the Australian Government National Health and Medical Research Council Standardised Participant Information and Consent Form for non-interventional studies, obtained from http://hrep.nhmrc.gov.au/toolbox/standardised-forms

Title: Validation of simulation as an assessment tool for assessing competence of medical students and junior doctors in managing cardiac emergencies

Supervisor: Professor Sandra Carr
Faculty of Medicine, Dentistry and Health Sciences
University of Western Australia

Co-Supervisor: Associate Professor William van Gaal
Director of Cardiology, The Northern Hospital
University of Melbourne

Student Investigator: Dr. Anastasia Vlachadis Castles
Masters student, University of Western Australia
The Northern Hospital

Location: The Northern Hospital, Epping, Victoria

Organisations Involved: University of Western Australia
Northern Health
The University of Melbourne

Part 1 What does participation involve?

1 Introduction

You are invited to take part in this research project, which is called Validation of simulation as an assessment tool for assessing competence of medical students and junior doctors in managing cardiac emergencies. You have been invited because as a medical student or junior doctor, you are an important target group for teaching and assessing management of cardiac emergencies. Your contact details were obtained from the University of Melbourne Northern Clinical School and/or Northern Health Junior Medical Workforce Unit.

This Participant Information Sheet/Consent Form tells you about the research project. It explains the processes involved with taking part. Knowing what is involved will help you decide if you want to take part in the research.

Please read this information carefully. Ask questions about anything that you don’t understand or want to know more about. Before deciding whether or not to take part, you might want to talk about it with a relative or friend.

Participation in this research is voluntary. If you don’t wish to take part, you don’t have to.
If you decide you want to take part in the research project, you will be asked to sign the consent section. By signing it you are telling us that you:

- Understand what you have read
- Consent to take part in the research project
- Consent to the use of your personal information as described.

You will be given a copy of this Participant Information and Consent Form to keep.

2 What is the purpose of this research?

From the first day of internship, junior medical officers are expected to be able to manage and lead cardiac (heart-related) emergencies with confidence and efficiency. Given that this is such a necessary skill, we need to be able to assess students and junior doctors’ readiness to manage cardiac emergencies. Simulation is a potential method to assess management of cardiac emergencies for medical students and junior doctors. However, there is currently insufficient evidence to support using simulation for this purpose.

The focal area of this proposed study is the validation of the assessment tools that will be utilised in the simulations. The results of this research have the potential to improve the quality of assessment and training of medical students and junior doctors in the area of cardiac emergencies.

The results of this research will be used by the researcher, Dr Anastasia Vlachadis Castles, to obtain a Master of Health Professional Education degree. This research is being conducted by University of Western Australia, in conjunction with Northern Health and the Northern Clinical School.

3 What does participation in this research involve?

A consent form will need to be signed prior to participants taking part in this research.

1. Participants will be asked to complete a questionnaire to gather information on demographics, education and experience.
2. Prior to participation in the educational program, participants will undertake a multiple choice question (MCQ) test, to be completed in thirty minutes.
3. Participants will engage in a structured simulated learning experience and assessment on the topic of cardiac emergency management. Participants will individually take part in one of two alternate simulated cardiac emergency scenarios. Debriefs will be conducted after each simulation by the assessors.

The total duration of participation is expected to be less than one hour per participant. There will be no follow-up required after the period of participation described above.

Participants will be required to attend The Northern Hospital and/or Northern Clinical School in order to participate. There are no direct costs associated with participating in this research project (you may participate on days when you usually attend the Hospital/Clinical School), nor will you be paid.

This research project has been designed to make sure the researchers interpret the results in a fair and appropriate way and avoids study doctors or participants jumping to conclusions.

4 Other relevant information about the research project

The expected number of participants in this research project is anywhere between 80-150. Participants will include Medical Students from Northern Clinical School and Junior Medical Officers (JMOs) employed by Northern Health. This is a cohort study so all participants will engage in an equivalent educational program. There is no control group. The Northern Health Clinical School has approved the involvement of their medical students in this project.
5 Do I have to take part in this research project?

Participation in any research project is voluntary. If you do not wish to take part, you do not have to. If you decide to take part and later change your mind, you are free to withdraw from the project at any stage.

If you do decide to take part, you will be given this Participant Information and Consent Form to sign and you will be given a copy to keep.

Your decision whether to take part or not to take part, or to take part and then withdraw, will not affect your education/training/employment, your relationship with professional staff or your relationship with The University of Melbourne, Northern Clinical School or Northern Health.

6 What are the possible benefits of taking part?

We cannot guarantee or promise that you will receive any benefits from this research; however, possible benefits may include: feedback on your performance in simulated cardiac emergencies, learning how to better manage cardiac emergencies.

7 What are the possible risks and disadvantages of taking part?

Psychological Distress: You may feel that some of the simulated situations are stressful or upsetting. If you do not wish to continue to participate you may stop immediately. If you become upset or distressed as a result of your participation in the research project, the research team will be able to arrange for counselling or other appropriate support free of charge.

8 What if I withdraw from this research project?

If you consent to participate, you may withdraw at any time. If you decide to withdraw from the project, please notify a member of the research team before you withdraw. If you do withdraw, you will be asked to complete and sign a ‘Withdrawal of Consent’ form; this will be provided to you by the research team and will need to be returned to Dr Anastasia Vlachadis Castles. Should the participants withdraw all the data will be destroyed unless otherwise agreed.

9 Could this research project be stopped unexpectedly?

This research project may be stopped unexpectedly for reasons such as illness/misadventure affecting members of the research team, insufficient number of participants or unexpected rates of psychological distress experienced by participants.

10 What happens when the research project ends?

Follow-up will not be undertaken beyond the one day of participation in the educational program/assessment. Participants will be provided with a summary of the results when the research project is completed (expected completion by end of 2015) through the Northern Clinical School or Northern Health Junior Medical Workforce Unit.

Part 2 How is the research project being conducted?

11 What will happen to information about me?

By signing the consent form you consent to the research team collecting and using personal information about you for the research project. Any information obtained in connection with this research project that can identify you will remain confidential. Your information will be used for the purpose of this research project and it will only be disclosed with your permission, except as required by law. Information obtained may also be used for further related research, including investigation of how demographic/education/experience factors affect participant performance in simulated cardiac emergencies and/or performance on written assessments.
Appendix 1: Page 4 of 7

The personal information that the research team collect and use will include questionnaires, multiple choice question test results and scores on simulations. Data will be de-identified at the earliest possible stage.

Data management will comply with The Australian Code for the Responsible Conduct of Research. Clear and accurate records will be kept documenting all stages of the research process. All data, hard copy and electronic, will be stored securely and only accessed by those involved in the research. All records will be retained for a minimum of 7 years after publication or project conclusion, whichever is later, and then be destroyed.

It is anticipated that the results of this research project will be published and/or presented in a variety of forums. In any publication and/or presentation, information will be provided in such a way that you cannot be identified, except with your express permission. Only group data will be published.

In accordance with relevant Australian and/or Victorian privacy and other relevant laws, you have the right to request access to the information about you that is collected and stored by the research team. You also have the right to request that any information with which you disagree be corrected. Please inform the research team member named at the end of this document if you would like to access your information.

12 Complaints and compensation

The researchers do not anticipate any loss or injury to participants. If you suffer any distress or psychological injury as a result of this research project, you should contact the research team as soon as possible. You will be assisted with arranging appropriate treatment and support.

Any complaints relating to your participation in this research should be directed to the research team and will be handled by Dr. Anastasia Vlachadis Castles and/or one of the senior researchers.

13 Who is organising and funding the research?

This research project is being conducted by Dr. Anastasia Vlachadis Castles, through the University of Western Australia.

The University of Western Australia, Northern Health and University of Melbourne may benefit financially from this research project if, for example, the project assists the institutions in any commercial enterprise.

You will not benefit financially from your involvement in this research project even if, for example, knowledge acquired from your information proves to be of commercial value to the institutions involved.

In addition, if knowledge acquired through this research leads to discoveries that are of commercial value to The University of Western Australia, Northern Health, University of Melbourne, the researchers or their institutions, there will be no financial benefit to you or your family from these discoveries.

No member of the research team will receive a personal financial benefit from your involvement in this research project (other than their ordinary wages).

14 Who has reviewed the research project?

All research in Australia involving humans is reviewed by an independent group of people called a Human Research Ethics Committee (HREC). The ethical aspects of this research project have been approved by the HRECs of the University of Western Australia and Northern Health, and registered with the HREC of the University of Melbourne. This project will be carried out according to the National Statement on Ethical Conduct in Human Research (2007). This statement has been developed to protect the interests of people who agree to participate in human research studies.
Appendix 1: Page 5 of 7

15 Further information and who to contact

The person you may need to contact will depend on the nature of your query. If you want any further information concerning this project or if you have any problems which may be related to your involvement in the project, or any complaints, you can contact the researcher at anastasia.castles@research.uwa.edu.au.

Research contact person

<table>
<thead>
<tr>
<th>Name</th>
<th>Dr. Anastasia Vlachadis Castles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Student researcher</td>
</tr>
<tr>
<td>Telephone</td>
<td>(03) 8405 8000</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:anastasia.castles@research.uwa.edu.au">anastasia.castles@research.uwa.edu.au</a> (email contact preferred)</td>
</tr>
</tbody>
</table>

If you have any complaints about any aspect of the project, the way it is being conducted or any questions about being a research participant in general, then you may contact the the Northern Health Research Governance Office at the Northern Health on (03) 8405 2918 or by emailing to ethics@nh.org.au

<table>
<thead>
<tr>
<th>Reviewing HREC name</th>
<th>University of Western Australia HREC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HREC Executive Officer</td>
<td>[Name]</td>
</tr>
<tr>
<td>Telephone</td>
<td>[HREC Executive Officer Phone number]</td>
</tr>
<tr>
<td>Email</td>
<td>[HREC Executive Officer Email address]</td>
</tr>
</tbody>
</table>

Approval to conduct this research has been provided by the University of Western Australia, Northern Health and University of Melbourne, in accordance with their ethics review and approval procedures. Any person considering participation in this research project, or agreeing to participate, may raise any questions or issues with the researchers at any time.

In addition, any person not satisfied with the response of researchers may raise ethics issues or concerns, and may make any complaints about this research project by contacting the Human Ethics Office at the University of Western Australia on (08) 6488 3703 or by emailing to humanethics@uwa.edu.au or they may contact the Northern Health Research Governance Office at the Northern Health on (03) 8405 2918 or by emailing to ethics@nh.org.au

All research participants are entitled to retain a copy of any Participant Information Form and/or Participant Consent Form relating to this research project.
Appendix 1: Page 6 of 7

Consent Form

Title: Validation of simulation as an assessment tool for assessing competence of medical students and junior doctors in managing cardiac emergencies

Supervisor: Professor Sandra Carr  
Faculty of Medicine, Dentistry and Health Sciences  
University of Western Australia

Co-Supervisor: Associate Professor William van Gaal  
Director of Cardiology, The Northern Hospital  
University of Melbourne

Student Investigator: Dr. Anastasia Vlachadis Castles  
Masters student, University of Western Australia  
The Northern Hospital

Location: The Northern Hospital, Epping, Victoria

Organisations Involved: University of Western Australia  
Northern Health  
The University of Melbourne

Declaration by Participant
I have read the Participant Information Sheet or someone has read it to me in a language that I understand. I understand the purposes, procedures and risks of the research described in the project. I have had an opportunity to ask questions and I am satisfied with the answers I have received. I freely agree to participate in this research project as described and understand that I am free to withdraw at any time during the project without affecting my future care and prejudice, in which case all the data will be destroyed unless otherwise agreed. I understand that I will be given a signed copy of this document to keep.

Name of Participant

Signature __________________________ Date __________________________

Declaration by Researcher*
I have given a verbal explanation of the research project, its procedures and risks and I believe that the participant has understood that explanation.

Name of Researcher*

Signature __________________________ Date __________________________

* An appropriately qualified member of the research team must provide the explanation of, and information concerning, the research project.

Note: All parties signing the consent section must date their own signature.
Form for Withdrawal of Participation

Title: Validation of simulation as an assessment tool for assessing competence of medical students and junior doctors in managing cardiac emergencies

Supervisor: Professor Sandra Carr
Faculty of Medicine, Dentistry and Health Sciences
University of Western Australia

Co-Supervisor: Associate Professor William van Gaal
Director of Cardiology, The Northern Hospital
University of Melbourne

Student Investigator: Dr. Anastasia Vlachadis Castles
Masters student, University of Western Australia
The Northern Hospital

Location: The Northern Hospital, Epping, Victoria

Organisations Involved: University of Western Australia
Northern Health
The University of Melbourne

Declaration by Participant
I wish to withdraw from participation in the above research project and understand that such withdrawal will not affect my routine care, or my relationships with the researchers or the institutions involved in the research.

Name of Participant

Signature ___________________________ Date ___________________________

In the event that the participant’s decision to withdraw is communicated verbally, the Senior Researcher must provide a description of the circumstances below.

Declaration by Researcher†
I have given a verbal explanation of the implications of withdrawal from the research project and I believe that the participant has understood that explanation.

Name of Researcher†

Signature ___________________________ Date ___________________________

† An appropriately qualified member of the research team must provide information concerning withdrawal from the research project.

Note: All parties signing the consent section must date their own signature.
Appendix 2: Participant Questionnaire

Appendix 2: Page 1 of 3

**PARTICIPANT IDENTIFICATION NUMBER (Researcher use only): __________________________**

**PARTICIPANT QUESTIONNAIRE**

Please write in block letters and place a tick in appropriate boxes ✅

**SECTION 1: DEMOGRAPHICS**

*All participants to complete this section*

1.1 **Surname:** __________________________
   
   **First name:** __________________________
   
1.2 **Age:** _________

1.3 **Gender:**
   
   - [ ] Female
   - [ ] Male

1.4 **Are you:**
   
   - [ ] Medical student ➔
     - [ ] MD2
     - [ ] MD3
     - [ ] MD4
   
   - [ ] JMO ➔
     - [ ] Intern
     - [ ] HM02- Medical stream
     - [ ] HM02- General stream
     - [ ] HM02- Surgical stream
     - [ ] HM02- Other (e.g. O&G/paeds)

**SECTION 2: MEDICAL EDUCATION**

*All participants to complete this section*

2.1 **Medical degree:**
   
   - [ ] University of Melbourne- graduate entry (MD)
   - [ ] University of Melbourne- graduate entry (MBBS- old course)
   - [ ] University of Melbourne- undergraduate entry (MBBS- old course)
   - [ ] Monash University- graduate entry
   - [ ] Monash University- undergraduate entry
   - [ ] Deakin University
   - [ ] Medical degree not listed above
     
     Please specify: __________________________
### Appendix 2: Page 2 of 3

<table>
<thead>
<tr>
<th>Section 3: Other Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>All participants to complete this section</td>
</tr>
</tbody>
</table>

3.1 Please specify **ANY** other degree(s) you **COMPLETED PRIOR** to enrolling in your medical degree:

---

3.2 Please specify **ANY** other degree(s) you **COMPLETED DURING OR AFTER** your medical degree:

---

3.3 Please specify **ANY** other degree(s) you **COMPLETED AT LEAST HALF OF** but do not intend to complete:

---

3.4 Please specify any other degree(s) you **ARE CURRENTLY ENROLLED IN** (apart from your medical degree if you are a current medical student):

---

### Section 4: Other Clinical Experience

All participants to complete this section

4.1 Have you previously worked in a hospital or healthcare setting in any other capacity **OTHER** than as a doctor? (eg. clerical, nurse, physiotherapist, occupational therapist, radiographer, etc)

- [ ] Yes, please specify:

- [ ] No

### Section 5: Clinical Experience - Medical Students

**Medical Students Only** to complete this section.

**Interns** please skip to Section 6, **HMO2 Participants** please skip to Section 7

5.1 Have you **completed**, or are you **currently completing**, an Emergency Medicine rotation:

- [ ] Yes
- [ ] No

5.2 Have you **completed**, or are you **currently completing**, a cardiology rotation?

- [ ] Yes
- [ ] No

*Medical students: You do not need to complete Sections 5 and 6. Thank you for completing this questionnaire.*
**Appendix 2: Page 3 of 3**

**SECTION 6: CLINICAL EXPERIENCE- INTERN**

*INTERNS ONLY to complete this section. HMO2 PARTICIPANTS please skip to Section 7*

6.1 Have you *completed*, or are you *currently completing*, your Emergency Medicine rotation:
- □ Yes
- □ No

6.2 How many medical rotations have you *ALREADY* completed this year?

Please include current rotation:
- □ 0
- □ 1
- □ 2

*Interns: You do not need to complete Section 7. Thank you for completing this questionnaire.*

**SECTION 7: CLINICAL EXPERIENCE- HMO2**

*HMO2 PARTICIPANTS ONLY to complete this section*

7.1 Have you *completed*, or are you *currently completing*, an Emergency Medicine rotation this year (HMO2)? This DOES NOT include an Emergency Medicine rotation completed during intern year.
- □ Yes
- □ No

7.2 How many medical rotations did you *complete* in your *INTERN* year?
- □ 1
- □ 2
- □ 3

7.3 How many Emergency Medicine rotations did you *complete* in your *INTERN* year?
- □ 1
- □ 2

7.4 Have you *completed*, or are you *currently completing*, a cardiology rotation?

- □ Yes, during internship
- □ Yes, during HMO2
- □ No

7.5 Please list rotations you have completed this year:

<table>
<thead>
<tr>
<th>Rotation</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Thank you for completing this questionnaire.*
Appendix 3: Multiple Choice Question Test

Appendix 3: Page 1 of 5

MULTIPLE CHOICE TEST

1. Segmental systolic dysfunction on transthoracic echocardiogram (TTE) suggests which underlying cause of heart failure?
   A. Hypertrophic cardiomyopathy
   B. Valvular disease
   C. Ischaemic heart disease

The following vignette relates to questions 2-4:
You are paged by nursing staff on Unit E to see Mr. Roberts, who is complaining of chest pain. You call the ward immediately and the nurse tells you that Mr. Roberts was admitted to hospital earlier the same day for an elective total knee replacement. He has not yet had his surgery. Five minutes ago he pressed the nurse call buzzer and told the nurse that he was experiencing central chest pain. She immediately performed a set of observations and asked another nurse to page you.

Mr. Roberts’ observations are as follows:
HR 100, BP 99/70, RR 22, SpO2 97% RA, temp 36.8

2. Which of the following do you ask the nurse to do while you are on your way to see the patient?
   A. Perform an ECG
   B. Give sublingual GTN
   C. Give oxygen via Hudson mask

From the end of the bed, the patient appears diaphoretic and quite pale, and has his hand is on the centre of his chest.
The patient’s ECG is shown below.

3. The culprit lesion is most likely in which vessel?
   A. Right coronary artery
   B. Left anterior descending artery
   C. Left circumflex artery
You ask the nurse to activate a MET call and Code STEMI.
A repeat set of observations is taken, which shows:
HR 120, BP 105/75, RR 25, SpO2 98%

The patient is complaining of central pain, which he reports is pressure-like and radiates to the left shoulder. He has not yet received any treatment for his chest pain.

4. The first step in treatment for this patient is to give:
A. ½ tab sublingual GTN + 300mg aspirin PO
B. ½ tab sublingual GTN + 2.5mg subcutaneous morphine
C. 300mg aspirin PO + 2.5mg subcutaneous morphine

5. Which of the following patients should proceed to revascularisation immediately?
A. A 56 year old male presenting with acute onset central chest pain. ECG shows ST elevation of ~2mm in leads V2, V3, V4. Initial troponin still pending.
B. A 45 year old female presenting with acute onset central chest pressure. ECG shows T wave inversion in leads II, III and aVF. Initial troponin is 0.90.
C. A 71 year old male presents with acute onset dyspnoea. ECG shows ST depression of ~1mm in leads V5-V6. Initial troponin is 1.05.

6. In addition to aspirin, which combination of medications should be used initially to manage ongoing pain in acute myocardial infarction?
A. Morphine, beta-blocker
B. GTN, morphine
C. Beta-blocker, GTN

The following vignette relates to questions 7-10:
Mrs. Jones, 47 year old lady, has just been brought in by ambulance to the emergency department. She presents with three episodes of shortness of breath over the past two days. She called an ambulance after experiencing a more severe episode while walking to work. She reports worsening shortness of breath on exertion over the last six months, and denies any chest pain, orthopnoea or paroxysmal nocturnal dyspnoea. Her past medical history includes rheumatoid arthritis, hypertension, diabetes and osteoarthritis.

On examination, heart sounds are dual, chest is clear, JVP is not elevated and there is no peripheral oedema. Vital signs, including BSL, are within normal range.
ECG shows T wave inversion in leads V1-V4, but no other abnormality. No previous ECGs are available for comparison.
Chest x-ray is clear.
Blood tests requested on initial assessment (FBE, UeC, CRP, LFTs) are normal except for a troponin level of 0.80.

7. Which of the following medications should you give as a stat dose at this point?
A. Aspirin 300mg PO
B. Metoprolol 25mg PO
C. Morphine 2.5mg subcutaneous

8. Which dose of enoxaparin will you prescribe for anticoagulation?
A. 40mg BD subcutaneous
B. 1mg/kg daily subcutaneous
C. 1mg/kg BD subcutaneous
9. When should troponin levels be repeated for this patient?
A. Only if pain recurs
B. Two hours after initial troponin
C. Six hours after initial troponin

10. What is the most appropriate timeline for revascularization for this patient?
A. Immediate thrombolysis if PCI (percutaneous coronary intervention) facilities are more than 90 mins away
B. Cardiac catheterisation +/- PCI as an inpatient
C. Cardiac catheterisation +/- PCI as an outpatient

11. Which of the following family histories is a significant risk factor for ischaemic heart disease?
A. Father and mother both died of acute myocardial infarctions, father aged 65 and mother aged 70.
B. Sister had a myocardial infarction aged 40 and is still alive 10 years later.
C. Brother died of myocardial infarction aged 60.

12. What is the best initial investigation to differentiate between dyspnoea due to pneumonia and dyspnoea due to decompensated heart failure?
A. CXR
B. FBE
C. D-dimer

13. Which of these blood tests is most useful to confirm a diagnosis of heart failure?
A. UeC
B. D-dimer
C. BNP

14. In a young patient with a first presentation of acute heart failure, which of the following is the most useful initial cardiac investigation?
A. CT coronary angiogram
B. Transthoracic echocardiogram (TTE)
C. Stress ECG

15. A sixty year old gentleman presents to the emergency department with central chest pain. His ECG shows ST elevation in leads V1-V4. He is diagnosed as having an ST elevation myocardial infarction (STEMI). His hospital records show that he underwent percutaneous intervention (PCI) with a bare metal stent (BMS) inserted in the proximal left anterior descending (LAD) artery to treat a non-ST elevation myocardial infarction (NSTEMI) approximately three weeks ago.
What is most likely to have precipitated the current cardiac event?
A. Non-compliance with anti-platelet medications
B. Non-compliance with beta-blockers
C. Non-compliance with a cardiac rehabilitation physiotherapy regime

16. Which of the following medication classes does not provide mortality benefit when initiated in the days following acute myocardial infarction?
A. Beta-blockers
B. ACE inhibitors
C. Calcium channel blockers
17. Which of the following findings on ECG would be most consistent with ST elevation myocardial infarction (STEMI)?
A. ST elevation of 1mm in V1, V2, V3
B. ST elevation of 2mm in aVF, V1, V2
C. ST elevation of 2mm in leads II, III, aVF

18. Which of the following is an equivalent to ST elevation myocardial infarction (STEMI), and in the context of chest pain, requires urgent revascularization?
A. ST depression in II, III, aVF
B. New 2nd degree AV block
C. New LBBB

19. Which of the following statements is correct?
A. Leads II, III, aVF correspond to the inferior territory, which is supplied by the right coronary artery (RCA) in most people
B. Leads V1-V4 correspond to the anteroseptal territory, which is supplied by the diagonal branch of the left circumflex artery in most people
C. Leads V5-V6, I and aVL correspond to the lateral territory, which is supplied by the left anterior descending (LAD) artery in most people

20. In managing a patient in acute pulmonary oedema, the patient should be positioned:
A. Supine
B. In the right decubitus position
C. Sitting up

21. In a patient with known heart failure who presents with acute decompensation of heart failure, which of the following additional measures is most appropriate?
A. Salt restriction
B. Fluid restriction
C. Intravenous fluids

22. In addition to myocardial infarction, common causes of acute pulmonary oedema, which need to be considered include:
A. Arrhythmia, renal failure
B. COPD, liver failure
C. COPD, anxiety

23. Which of the following is the LEAST IMPORTANT sign to assess in a patient with suspected acute pulmonary oedema:
A. Tachypnoea/dyspnoea
B. Pulsatile liver
C. Elevated JVP

24. Which of the following is the most important combination of vital sign parameters in monitoring the progress of a patient with acute pulmonary oedema?
A. Respiratory rate and oxygen saturations
B. Heart rate and respiratory rate
C. Oxygen saturations and temperature
25. Which of the following is NOT an initial investigation that should be carried out immediately in a patient with suspected acute pulmonary oedema:
A. ABG
B. CXR
C. Coronary angiogram

The following vignette relates to questions 26-30:
Mrs. Green, a 68 year old lady, is brought into the emergency department by ambulance with acute onset shortness of breath. She denies chest pain and nausea/vomiting but reports some ‘dizziness’.

Mrs. Green has a history of ischaemic heart disease, chronic heart failure, atrial fibrillation (AF), type 2 diabetes mellitus, hypertension and gastro-oesophageal reflux (GORD). Her regular medications include: aspirin 100mg daily, metoprolol 25mg BD, atorvastatin 80mg daily, frusemide 40mg BD, rivaroxaban 20mg daily, metformin 500mg BD, perindopril 4mg daily, pantoprazole 40mg daily

Mrs. Green has a heart rate of 130, blood pressure of 140/75, respiratory rate of 32, oxygen saturation of 91% on 6L/min oxygen via nasal prongs, and temperature of 36.8.
On examination, she has an irregularly irregular pulse, JVP of 6cm, crackles to the midzones bilaterally and pitting oedema to just above the knees.

26. Which of these is an appropriate initial dose of frusemide to give in this situation?
A. 20mg PO
B. 40mg IV
C. 80mg IV

27. ECG shows AF with rapid ventricular response (ventricular rate 140). Initial management for this patient should be:
A. 25mg metoprolol PO
B. Electrical cardioversion
C. Digoxin loading

28. Before any treatment for AF is given, the patient spontaneously reverts to sinus rhythm but continues to be very short of breath. Oxygen saturation is now 88% on 6L O2 via nasal prongs.
The most appropriate next step in management is:
A. Increasing oxygen to 10L/min via Hudson mask
B. Giving aspirin
C. Giving IV fluids

29. Mrs. Green’s respiratory status continues to worsen but she remains alert. Her ABG shows that she is hypoxic but not significantly hypercapnic or acidotic. You call the ICU registrar because you believe the patient may now require which of the following:
A. BiPAP (bi-level positive airway pressure)
B. CPAP (continuous positive airway pressure)
C. Intubation and IPPV (intermittent positive pressure ventilation)

30. Which of these is the best way to accurately monitor Mrs. Green’s fluid status?
A. Insertion of a urinary catheter, strict fluid balance
B. Daily weight
C. Twice daily clinical examination for fluid statusp
MULTIPLE CHOICE TEST - ANSWERS & FEEDBACK

1. Segmental systolic dysfunction on transthoracic echocardiogram (TTE) suggests which underlying cause of heart failure?
   A. Hypertrophic cardiomyopathy
   B. Valvular disease
   C. Ischaemic heart disease
   [answer: C]

   Explanation: Segmental systolic dysfunction on TTE describes dysfunction localised to a specific area of myocardium. This suggests ischaemic heart disease, as a territory (or territories) supplied by a diseased coronary artery (or arteries) is affected but there is not global dysfunction. Hypertrophic cardiomyopathy appears on TTE as areas of hypertrophied myocardium and generally affects the left ventricle. Valvular disease is demonstrated by abnormal flow of blood through a valve (or valves), and can cause dilatation or hypertrophy of isolated chambers, but this is distinct from segmental dysfunction.

   The following vignette relates to questions 2-4:
   You are paged by nursing staff on Unit E to see Mr. Roberts, who is complaining of chest pain. You call the ward immediately and the nurse tells you that Mr. Roberts was admitted to hospital earlier the same day for an elective total knee replacement. He has not yet had his surgery. Five minutes ago he pressed the nurse call buzzer and told the nurse that he was experiencing central chest pain. She immediately performed a set of observations and asked another nurse to page you.

   Mr. Roberts’ observations are as follows:
   HR 100, BP 99/70, RR 22, SpO2 97% RA, temp 36.8

2. Which of the following do you ask the nurse to do while you are on your way to see the patient?
   A. Perform an ECG
   B. Give sublingual GTN
   C. Give oxygen via Hudson mask
   [answer: A]

   Explanation: In the case of chest pain, an ECG is ALWAYS a necessary initial investigation. It is important to consider acute myocardial infarction as an important cause of chest pain. Sublingual GTN should be used with caution if the patient’s blood pressure is low. Supplemental oxygen has dubious utility when the patient’s oxygen saturation is normal.
From the end of the bed, the patient appears diaphoretic and quite pale, and has his hand is on the centre of his chest. The patient’s ECG is shown below.

3. The culprit lesion is most likely in which vessel?
A. Right coronary artery
B. Left anterior descending artery
C. Left circumflex artery
[answer: B]
Explanation: This ECG shows significant ST segment elevation mostly in leads V1-V4 - anteroseptal leads. This corresponds to left anterior descending (LAD) artery territory. The ST depression in the inferior leads (II, III, aVF) is a reciprocal change. Usually, a lesion in the right coronary artery would cause ST elevation in inferior leads (II, III, aVF), while a lesion in the left circumflex artery would cause ST elevation in lateral leads (V5, V6, I, aVL).

You ask the nurse to activate a MET call and Code STEMI. A repeat set of observations is taken, which shows: HR 120, BP 105/75, RR 25, SpO2 98%

The patient is complaining of central pain, which he reports is pressure-like and radiates to the left shoulder. He has not yet received any treatment for his chest pain.

4. The first step in treatment for this patient is to give:
A. ½ tab sublingual GTN + 300mg aspirin PO
B. ½ tab sublingual GTN + 2.5mg subcutaneous morphine
C. 300mg aspirin PO + 2.5mg subcutaneous morphine
[answer: A]
Explanation: GTN and aspirin should be used in the initial treatment of acute myocardial infarction. A loading dose of aspirin (300mg) is used as an antiplatelet agent. GTN is used to dilate the coronary arteries. GTN causes blood pressure to decrease so should be used with caution when blood pressure is low. In this case the blood pressure is now sufficient for GTN to be given. Morphine is generally used to control pain as a second line after GTN.

5. Which of the following patients should proceed to revascularisation immediately?
A. A 56 year old male presenting with acute onset central chest pain. ECG shows ST elevation of ~2mm in leads V2, V3, V4. Initial troponin still pending.
B. A 45 year old female presenting with acute onset central chest pressure. ECG shows T wave inversion in leads II, III and aVF. Initial troponin is 0.90.
C. A 71 year old female presents with acute onset dyspnoea. ECG shows ST depression of ~1mm in leads V5-V6. Initial troponin is 1.05.
[answer: A]
Explanation: In answer A, the diagnosis is ST-elevation myocardial infarction (STEMI). This is an indication for immediate revascularisation. In answers B and C, the diagnosis is non-ST-elevation myocardial infarction (NSTEMI), and these patients should undergo coronary angiography within 48hrs, unless the risk of coronary angiography exceeds the potential benefit.
6. In addition to aspirin, which combination of medications should be used initially to manage ongoing pain in acute myocardial infarction?
A. Morphine, beta-blocker
B. GTN, morphine
C. Beta-blocker, GTN
[answer: B]

Explanations: GTN is indicated for acute management of acute myocardial infarction. Morphine is used if pain does not resolve with GTN. Beta-blockers are indicated for longer-term medical management of ischaemic heart disease but should be used with caution in initial management, especially where heart failure or arrhythmias are present.

The following vignette relates to questions 7-10:
Mrs. Jones, 47 year old lady, has just been brought in by ambulance to the emergency department. She presents with three episodes of shortness of breath over the past two days. She called an ambulance after experiencing a more severe episode while walking to work. She reports worsening shortness of breath on exertion over the last six months, and denies any chest pain, orthopnoea or paroxysmal nocturnal dyspnoea. Her past medical history includes rheumatoid arthritis, hypertension, diabetes and osteoarthritis.

On examination, heart sounds are dual, chest is clear, JVP is not elevated and there is no peripheral oedema. Vital signs, including BSL, are within normal range. ECG shows T wave inversion in leads V1-V4, but no other abnormality. No previous ECGs are available for comparison.

Chest x-ray is clear.
Blood tests requested on initial assessment (FBE, UeC, CRP, LFTs) are normal except for a troponin level of 0.80.

7. Which of the following medications should you give as a stat dose at this point?
A. Aspirin 300mg PO
B. Metoprolol 25mg PO
C. Morphine 2.5mg subcutaneous
[answer: A]

Explanations: This patient is having a non-ST-elevation myocardial infarction (NSTEMI). As previously mentioned, aspirin is used in the initial treatment of acute myocardial infarction. There is no indication for metoprolol at this point. Morphine is not necessary as there is no pain currently, and other measures should be taken before considering using morphine to treat dyspnoea.

8. Which dose of enoxaparin will you prescribe for anticoagulation?
A. 40mg BD subcutaneous
B. 1mg/kg daily subcutaneous
C. 1mg/kg BD subcutaneous
[answer: C]

Explanations: The therapeutic dose of enoxaparin, used to treat acute myocardial infarction where the patient is not undergoing primary PCI, is 1mg/kg BD. Dosage is adjusted (decreased) for renal impairment.

9. When should troponin levels be repeated for this patient?
A. Only if pain recurs
B. Two hours after initial troponin
C. Six hours after initial troponin
[answer: C]

Explanations: Troponin levels correlate with prognosis. In patients not undergoing revascularisation, troponin levels should be repeated every six hours until troponin is seen to peak.
10. What is the most appropriate timeline for revascularization for this patient?
A. Immediate thrombolysis if PCI (percutaneous coronary intervention) facilities are more than 90mins away
B. Cardiac catheterisation +/- PCI as an inpatient
C. Cardiac catheterisation +/- PCI as an outpatient
[answer: B]
Explanation: According to Australian guidelines, a patient who has had a non-ST-elevation myocardial infarction (NSTEMI) should undergo coronary angiography within 48hrs, unless the risk of coronary angiography exceeds the potential benefit.

11. Which of the following family histories is a significant risk factor for ischaemic heart disease?
A. Father and mother both died of acute myocardial infarctions. father aged 65 and mother aged 70.
B. Sister had a myocardial infarction aged 40 and is still alive 10 years later.
C. Brother died of myocardial infarction aged 60.
[answer: B]
Explanation: “A positive family history includes any first-degree male relative aged 45 years or younger or any first-degree female relative aged 55 years or younger who experienced a myocardial infarction.” (Reference: Medscape)

12. What is the best initial investigation to differentiate between dyspnoea due to pneumonia and dyspnoea due to decompensated heart failure?
A. CXR
B. FBE
C. D-dimer
[answer: A]
Explanation: Chest x-ray findings are different in pneumonia and decompensated heart failure. In pneumonia, consolidation, often of an isolated area of lung, is seen. In decompensated heart failure, pulmonary congestion, upper lobe diversion, and sometimes pleural effusion, is seen. White cell count will usually be elevated in pneumonia but this is non-specific. D-dimer is used to exclude pulmonary embolism (PE) in patients with low pre-test probability of PE.

13. Which of these blood tests is most useful to confirm a diagnosis of heart failure?
A. UeC
B. D-dimer
C. BNP
[answer: C]
Explanation: Renal function may be affected in heart failure but this is neither sensitive nor specific. D-dimer is used to exclude pulmonary embolism (PE) in patients with low pre-test probability of PE. BNP is elevated in heart failure with high sensitivity and relatively high specificity, depending on the cut off level used.
14. In a young patient with a first presentation of acute heart failure, which of the following is the most useful initial cardiac investigation?
A. CT coronary angiogram
B. Transthoracic echocardiogram (TTE)
C. Stress ECG
[answer: B]
Explanation: A CT coronary angiogram is used to assess patients at intermediate risk of ischaemic heart disease and gives predominantly structural information on the coronary arteries (and, depending on technique used, left ventricle). TTE gives functional information and can demonstrate both systolic and diastolic dysfunction, therefore also differentiating between heart failure with reduced ejection fraction (previously called systolic heart failure) and heart failure with preserved ejection fraction (previously called diastolic heart failure). Stress ECG may show evidence of myocardial ischaemia on exertion or arrhythmia, but is not useful in assessing heart failure.

15. A sixty year old gentleman presents to the emergency department with central chest pain. His ECG shows ST elevation in leads V1-V4. He is diagnosed as having an ST elevation myocardial infarction (STEMI). His hospital records show that he underwent percutaneous intervention (PCI) with a bare metal stent (BMS) inserted in the proximal left anterior descending (LAD) artery to treat a non-ST elevation myocardial infarction (NSTEMI) approximately three weeks ago.
What is most likely to have precipitated the current cardiac event?
A. Non-compliance with anti-platelet medications
B. Non-compliance with beta-blockers
C. Non-compliance with a cardiac rehabilitation physiotherapy regime
[answer: A]
Explanation: Non-compliance with anti-platelet medications, especially within the first year post angioplasty, is most likely to cause in-stent thrombosis. Non-compliance with beta-blockers or cardiac rehabilitation physiotherapy is more likely to cause poorer outcomes in the medium- to long-term.

16. Which of the following medication classes does not provide mortality benefit when initiated in the days following acute myocardial infarction?
A. Beta-blockers
B. ACE inhibitors
C. Calcium channel blockers
[answer: C]
Explanation: “Treatment with both beta-adrenergic blockers and ACE inhibitors may improve the balance between myocardial oxygen supply and demand. And it may limit infarct size.” “Calcium channel blockers have not been beneficial in acute myocardial infarction, and they may exert deleterious adverse effects alone or when given with other medications. Therefore, they should generally be avoided.”
(Reference: Medscape)

17. Which of the following findings on ECG would be most consistent with ST elevation myocardial infarction (STEMI)?
A. ST elevation of 1mm in V1, V2, V3
B. ST elevation of 2mm in aVF, V1, V2
C. ST elevation of 2mm in leads II, III, aVF
[answer: C]
Explanation: Diagnosis of STEMI requires ST elevation of ≥2mm in males, or 1.5mm in females, in two contiguous precordial leads; or ≥1mm ST elevation in two other contiguous leads.
18. Which of the following is an equivalent to ST elevation myocardial infarction (STEMI), and in the context of chest pain, requires urgent revascularization?
A. ST depression in II, III, aVF
B. New 2nd degree AV block
C. New LBBB
[answer: C]
Explanation: New LBBB is an equivalent to ECG finding to STEMI and requires urgent revascularisation.

19. Which of the following statements is correct?
A. Leads II, III, aVF correspond to the inferior territory, which is supplied by the right coronary artery (RCA) in most people
B. Leads V1-V4 correspond to the anteroseptal territory, which is supplied by the diagonal branch of the left circumflex artery in most people
C. Leads V5-V6, I and aVL correspond to the lateral territory, which is supplied by the left anterior descending (LAD) artery in most people
[answer: A]
Explanation: Leads II, III, aVF corresponds to inferior territory, supplied by the RCA. Leads V1-V4 corresponds to anteroseptal territory, which is supplied by the LAD in most people. Leads V5-V6, I and aVL corresponds to lateral territory, which is supplied by the left circumflex.

20. In managing a patient in acute pulmonary oedema, the patient should be positioned:
A. Supine
B. In the right decubitus position
C. Sitting up
[answer: C]
Explanation: Patients with acute pulmonary oedema generally experience orthopnoea, and dyspnoea is improved with sitting as upright as possible.

21. In a patient with known heart failure who presents with acute decompensation of heart failure, which of the following additional measures is most appropriate?
A. Salt restriction
B. Fluid restriction
C. Intravenous fluids
[answer: B]
Explanation: Intravenous fluids will worsen acute decompensated heart failure and should only be given if the patient is very hypotensive, and even hypotensive, with caution. Fluid restriction is important to ensure that fluid overload is not exacerbated, and to minimise further strain being put on the heart. Salt restriction may be used in heart failure due to the effect of salt on fluid retention, but is only useful for longer-term management.

22. In addition to myocardial infarction, common causes of acute pulmonary oedema, which need to be considered include:
A. Arrhythmia, renal failure
B. COPD, liver failure
C. COPD, anxiety
[answer: A]
Explanation: Usually acute pulmonary oedema is precipitated by an acute pathology. Acute myocardial infarction, arrhythmias (eg. atrial fibrillation) and renal failure (acute or acute on chronic) are known precipitants of acute pulmonary oedema, especially in those with underlying heart failure.
23. Which of the following is the LEAST IMPORTANT sign to assess in a patient with suspected acute pulmonary oedema:
A. Tachypnoea/dyspnoea
B. Pulsatile liver
C. Elevated JVP
[answer: B]
Explaination: Tachypnoea/dyspnoea and elevated JVP are expected acute findings in pulmonary oedema. A pulsatile liver is usually seen in tricuspid regurgitation, which is usually chronic.

24. Which of the following is the most important combination of vital sign parameters in monitoring the progress of a patient with acute pulmonary oedema?
A. Respiratory rate and oxygen saturations
B. Heart rate and respiratory rate
C. Oxygen saturations and temperature
[answer: A]
Explaination: In acute pulmonary oedema, patients become dyspnoeic and, therefore, tachypnoeic. Oxygen saturations tend to be decreased in pulmonary oedema, due to impaired gas exchange. Heart rate will often be elevated but will not reliably change with treatment. Temperature will not usually be elevated unless there is an underlying infectious cause for the pulmonary oedema.

25. Which of the following is NOT an initial investigation that should be carried out immediately in a patient with suspected acute pulmonary oedema:
A. ABG
B. CXR
C. Coronary angiogram
[answer: C]
Explaination: ABG should be performed to assist in assessing the degree of impairment of gas exchange, and to assess for acid-base disturbance. Chest x-ray can be used to differentiate between different causes of dyspnoea (eg. pulmonary oedema vs pneumonia) and assess the extent/severity of pulmonary oedema. Coronary angiography is not necessary unless indicated (eg. suspected myocardial infarction as the underlying cause of pulmonary oedema).
Appendix 4: Page 8 of 9

The following vignette relates to questions 26-30:

Mrs. Green, a 68 year old lady, is brought into the emergency department by ambulance with acute onset shortness of breath. She denies chest pain and nausea/vomiting but reports some ‘dizziness’.

Mrs. Green has a history of ischaemic heart disease, chronic heart failure, atrial fibrillation (AF), type 2 diabetes mellitus, hypertension and gastro-oesophageal reflux (GORD). Her regular medications include: aspirin 100mg daily, metoprolol 25mg BD, atorvastatin 80mg daily, frusemide 40mg BD, rivaroxaban 20mg daily, metformin 500mg BD, perindopril 4mg daily, pantoprazole 40mg daily.

Mrs. Green has a heart rate of 130, blood pressure of 140/75, respiratory rate of 32, oxygen saturation of 91% on 6L/min oxygen via nasal prongs, and temperature of 36.8.

On examination, she has an irregularly irregular pulse, JVP of 6cm, crackles to the midzones bilaterally and pitting oedema to just above the knees.

26. Which of these is an appropriate initial dose of frusemide to give in this situation?
   A. 20mg PO
   B. 40mg IV
   C. 80mg IV
   [answer: B]

   Explanation: In acute pulmonary oedema, oral frusemide will act too slowly, so IV frusemide is required. In treating acute pulmonary oedema, the patient's total daily frusemide dose is usually given for initial management. 40mg IV frusemide is equivalent in effect to 80mg oral frusemide.

27. ECG shows AF with rapid ventricular response (ventricular rate 140). Initial management for this patient should be:
   A. 25mg metoprolol PO
   B. Electrical cardioversion
   C. Digoxin loading
   [answer: A]

   Explanation: Given that the patient is known to have atrial fibrillation and is already on metoprolol, the issue is this case is that the patient is not currently rate-controlled. As such, metoprolol, starting at 25mg, should be used for initial management. Digoxin loading may be considered if rate cannot be controlled with metoprolol (or other beta blockers). Electrical cardioversion may be considered for persistent AF, or new AF, and is usually performed as an elective procedure.

28. Before any treatment for AF is given, the patient spontaneously reverts to sinus rhythm but continues to be very short of breath. Oxygen saturation is now 88% on 6L O2 via nasal prongs.

   The most appropriate next step in management is:
   A. Increasing oxygen to 10L/min via Hudson mask
   B. Giving aspirin
   C. Giving IV fluids
   [answer: A]

   Explanation: Initial management of acute pulmonary oedema includes frusemide, morphine, nitrates (eg. GTN) and oxygen. Non-invasive ventilation (CPAP) may also be required. There is no indication to give aspirin in this case. Giving IV fluids in a patient who is already fluid overloaded would be detrimental.
29. Mrs. Green's respiratory status continues to worsen but she remains alert. Her ABG shows that she is hypoxic but not significantly hypercapnic or acidotic. You call the ICU registrar because you believe the patient may now require which of the following:
   A. BiPAP (bi-level positive airway pressure)
   B. CPAP (continuous positive airway pressure)
   C. Intubation and IPPV (intermittent positive pressure ventilation)
   [answer: B]
   *Explanation: CPAP is used in acute pulmonary oedema where there is insufficient improvement with medications and oxygen.*

30. Which of these is the best way to accurately monitor Mrs. Green's fluid status?
   A. Insertion of a urinary catheter, strict fluid balance
   B. Daily weight
   C. Twice daily clinical examination for fluid status
   [answer: A]
   *Explanation: Where accurate assessment of fluid status is required, it is important to measure both fluid intake and fluid output (i.e., urine output). This is achieved by keeping a strict fluid balance, which includes measuring urine output accurately, which is best achieved by using a urinary catheter. Daily weight and clinical examination are also crucial components of monitoring fluid status but have inadequate accuracy for determining change in fluid status; weight can be affected by other factors (e.g., different scales, timing, irregular bowel motions) and clinical examination is somewhat subjective.*
## Appendix 5: Simulation Scenarios

### Title:
Validation of simulation as an assessment tool for assessing competence of medical students and junior doctors in managing cardiac emergencies

### Supervisor:
Professor Sandra Carr  
Faculty of Medicine, Dentistry and Health Sciences  
University of Western Australia

### Co-Supervisor:
Associate Professor William van Gaal  
Director of Cardiology, The Northern Hospital  
University of Melbourne

### Student Investigator:
Dr. Anastasia Vlachadis Castles  
Masters student, University of Western Australia  
The Northern Hospital

### Location:
The Northern Hospital, Epping, Victoria

### Organisations Involved:
University of Western Australia  
Northern Health  
The University of Melbourne
TABLE OF CONTENTS

SIMULATION SCENARIO 1: ACUTE MYOCARDIAL INFARCTION

I. TARGET AUDIENCE
II. LEARNING OUTCOMES
III. CASE OVERVIEW
IV. ENVIRONMENT
V. ACTORS
VI. CASE DETAILS
VII. DEBRIEF

SIMULATION SCENARIO 2: ACUTE PULMONARY OEDEMA

I. TARGET AUDIENCE
II. LEARNING OUTCOMES
III. CASE OVERVIEW
IV. ENVIRONMENT
V. ACTORS
VI. CASE DETAILS
VII. DEBRIEF

APPENDICES

Appendix 1.1: Triage Notes
Appendix 1.2: Vital Signs
Appendix 1.3: ECG #1
Appendix 1.4: ABG/VBG
Appendix 1.5: Simulation Assessment Sheet

Appendix 2.1: Patient Notes
Appendix 2.2: Vital Signs
Appendix 2.3: ECG #2
Appendix 2.4: ABG/VBG Results
Appendix 2.5: Simulation Assessment Sheet
SIMULATION SCENARIO 1: ACUTE MYOCARDIAL INFARCTION

I. TARGET AUDIENCE
   - University of Melbourne Medical Students- Years 2-4
   - Northern Health Junior Medical Officers- Postgraduate years 1-2

II. LEARNING OUTCOMES
    (Based on Australian Curriculum Framework for Junior Doctors)

A. Patient assessment
   1. History
      - takes a focused history to differentiate between causes of chest pain
   2. Examination
      - performs a focused cardiorespiratory examination
   3. Investigation
      - requests appropriate investigations
      - interprets an ECG
   4. Prioritisation
      - recognises patients who are acutely ill and require immediate treatment
      - recognises when to call for help
   5. Diagnosis
      - is able to form a list of differential diagnoses and determine the most likely diagnosis

B. Patient management
   1. Management
      - identifies and initiates appropriate initial management for acute coronary syndrome
      - identifies the need for senior clinician input to determine and implement further management
   2. Patient progress
      - re-evaluates the patient’s progress and response to treatment

III. CASE OVERVIEW
    This scenario depicts the case of a 62 year old man presenting to the emergency department (ED) with chest pain.

IV. ENVIRONMENT
A. Room set up
   - simulation laboratory at The Northern Hospital- ward set up
B. Patient set up
   - actor as patient
   - seated in bed
   - wearing patient gown
C. Props
   - triage notes with initial vital signs (Appendix 1.1: Vital Signs)
   - vital signs monitor (Results as per Appendix 1.2: Vital Signs)
   - cannulation trolley
   - printed ECG (Appendix 1.3: ECG #1- anterior STEMI)
   - crash cart (standard set up)-out of view, brought in if required
   - medications available (aspirin, anginine, GTN patch, morphine, frusemide, metoprolol, clopidogrel)
   - oxygen- nasal prongs & Hudson mask
   - printed ABG/VBG results (Appendix 1.4: ABG/VBG results)

V. ACTORS
A. Patient- 62 year old man/woman with chest pain
B. ED nurse
VI. CASE DETAILS
A. Initial information given to participant:
- You are a junior doctor working in the emergency department at The Northern Hospital.
- You have been asked to see Mr. Smith, a 62 year old man who presented to the emergency department with chest pain. He is triage category 2.

OR
- You are a junior doctor working in the emergency department at The Northern Hospital.
- You have been asked to see Mrs. Smith, a 62 year old woman who presented to the emergency department with chest pain. She is triage category 2.

B. Handover from ED nurse
- “I’ve just received Mr(x). Smith from triage. (S)he’s complaining of central chest pain and seems a bit short of breath. I’ve just got him/her changed into a gown and monitored.”

C. History from patient
1. Personal information
   - John/Joan Smith
   - 62 years old

2. HOPC
   * Volunteer the following information only:
     - severe chest pain
     - “the pain wasn’t going away, so I drove myself to the hospital”
   * Give the following information on direct questioning:
     * chest pain
       - onset: sudden, approximately an hour ago
       - location: pain over lower left side of chest
       - radiation: radiates up to left side of jaw
       - NO radiation to back/shoulders/arms
       - character: “like somebody is pushing on my chest”
       - constant
       - NOT pleuritic
       - severity: initially 5/10, now 9/10
       - aggravating/relieving factors
         - none
         - “I tried taking Mylanta because sometimes I get heartburn, but it didn’t work”
       - previous episodes
         - “I sometimes get chest pain when I’m playing tennis but never like this”
         - “just over the last few months”
       - exercise tolerance
         - “I go for walks with my partner a couple of nights per week. We walk for about half an hour”
     * associated symptoms
       - dyspnoea
       - nausea, 1 x vomiting after onset of chest pain
       - “felt a bit hot”
       - NO lightheadedness
       - NO cough
       - NO leg pain
       - NO recent illness

* systems review- all negative
Appendix 5: Page 5 of 30

3. Risk factors
   - for IHD - HTN, FHx (father AMI aged 45, brother AMI aged 50), ex-smoker
   - unsure re: cholesterol
   - NO diabetes
   - for VTE - NO recent travel/surgery/immobility, NO personal Hx or FHx of VTE
   - for female pt- no HRT

4. PMHx
   - “I don’t go to the doctor much”
   - HTN - “for at least 10 years”
   - hypothyroidism - “diagnosed last year when I was feeling tired all the time”
   - appendicectomy aged 19

5. Medications
   - perindopril 5mg daily
   - thyroxine 50mcg daily

6. Allergies
   - penicillin- rash

7. Social Hx
   - home with partner
   - 2 adult children in Queensland
   - moved to Victorian from Queensland one year ago
   - accountant
   - ex-smoker- smoked 1 pack per day from age 15 to 55, then quit
   - ETOH- 2 glasses of wine per night, “sometimes a few more when I’m out”
   - denies drug use

8. FHx - father AMI aged 45, brother AMI aged 50

D. Examination findings (given by assessor as participant performs that part of examination)
   1. Appearance
      - appears anxious
   2. Vital signs
      - on monitor- see Appendix 1.2
   3. Neck- JVPNE
   4. Chest- not tender to palpation
      - apex beat not palpable, no heaves/thrills
      - HS dual, no added sounds
      - lungs clear, normal air entry
   5. Abdomen
      - SNT
      - appendicectomy scar RIF
   6. Legs - no peripheral oedema
      - calves SNT
   * No other positive examination findings

E. ECG
   Appendix 1.3- ECG #1- Anterior STEMI
F. Nurse instructions:
* ONLY perform tasks within your scope of practice
* When performing tasks, do so in real time (ie do not perform the task faster than you really could)
* Invasive tasks should be imitated only (eg, venepuncture, cannulation, administering medications)

* Below is a list of responses to questions/requests from participants:
1. Asked for vital signs
   - ask participant to refer to triage notes or monitor
2. Asked if the patient has had aspirin
   - direct participant to drug chart
   - if the participant asks for aspirin again, show where aspirin was given at triage
3. Asked to perform ECG
   - give participant ECG #1 (Appendix 1.3)
4. Asked to take bloods
   - ask for pathology slip, ask if the patient needs a cannula
5. Asked to insert a cannula
   - ask if the patient needs bloods taken too, ask for a pathology slip
6. Asked to arrange a chest x-ray
   - ask for request form, organise PSA to take request to radiology (return in 30 seconds)
7. Asked to give patient oxygen
   - ask how much; if Hudson mask not specified, use nasal prongs
8. Asked to give any medications
   - ensure they are charted first
   - question any unusual doses/routes of administration
9. Asked for previous records/old ECG
   - the patient has not been here before, he moved here from interstate one year ago
10. Asked to seek assistance of a senior doctor
    - return after 1 minute, inform participant that the senior doctor is with a very unwell patient but will come to assist as soon as possible

G. Scenario progress
1. Pain - will remain the same severity without treatment
   - 1st dose ¼ anginine S/L will not be effective
   - 2nd dose ½ anginine S/L will improve pain to 6/10 but cause headache
   - further doses of anginine S/L will improve pain by 1/10 per dose and worsen headache
   - morphine given after at least 2 doses of anginine will improve pain to 3/10
   - will be unaffected by oxygen
2. Dyspnœa
   - will improve slightly with oxygen
   - will not improve with frusemide
3. Nausea/vomiting
   - no further vomiting
   - nausea will improve with metoclopramide or ondansetron/similar
4. Vital signs - see Appendix 1.2: Vital Signs
5. ECG - same ECG (Appendix 1.3: ECG #1) if repeated
6. Blood tests
   - results will not be available by the end of the simulation
   - if VBG is requested at time of venepuncture/cannulation, results will be available in 3 minutes (Appendix 1.4: VBG)
   - if VBG requested after initial blood tests sent, repeat venepuncture will be required, results of VBG will be available 3 minutes after venepuncture (Appendix 1.4: VBG)
   - if the participant performs an ABG, results will be available after 3 minutes (Appendix 1.4: ABG)
7. CXR
   - will not be performed by the end of the simulation
8. Request for assistance by registrar/consultant
   - assistance will arrive at 9 minutes
VII. DEBRIEF
Five minutes total structured brief

A. Structure
1. Ask the participant what they think they did well
2. Tell the participant which aspects they performed well on
3. Ask the participant what they think they could improve on
4. Tell the participant which aspects they performed less well on
5. Ask the participant what they thought was happening in the scenario
6. Explain basic management of acute myocardial infarction, if required

B. General guidelines
1. Do not divulge the participant’s scores
2. Do not compare the participant’s performance to that of other participants
3. If the participant appears to be distressed at their performance or the simulation outcome, please direct them to the Student Investigator (Dr Anastasia Vlachadis Castles).
SIMULATION SCENARIO 2: ACUTE PULMONARY OEDEMA

I. TARGET AUDIENCE
   • University of Melbourne Medical Students- Years 2-4
   • Northern Health Junior Medical Officers- Postgraduate years 1-2

II. LEARNING OUTCOMES
    (Based on Australian Curriculum Framework for Junior Doctors)
    A. Patient assessment
       1. History
          - takes a focused history to differentiate between causes of dyspnoea
       2. Examination
          - performs a focused cardiorespiratory examination
       3. Investigation
          - requests appropriate investigations
          - interprets an ECG
       4. Prioritisation
          - recognises patients who are acutely ill and require immediate treatment
          - recognises when to call for help
       5. Diagnosis
          - is able to form a list of differential diagnoses and determine the most likely diagnosis
    B. Patient management
       1. Management
          - identifies and initiates appropriate initial management for acute pulmonary oedema/
            fluid overload
          - identifies the need for senior clinician input to determine and implement further management
       2. Patient progress
          - re-evaluates the patient’s progress and response to treatment

III. CASE OVERVIEW
    This scenario depicts the case of a 76 year old male, an inpatient admitted with cellulitis, now complaining of dyspnoea.

IV. ENVIRONMENT
    A. Room set up
       - simulation laboratory at The Northern Hospital- ward set up
    B. Patient set up
       - actor as patient
       - seated in bed
       - wearing patient gown
       - IV fluids running
    C. Props
       - patient notes folder (Appendix 2.1: Patient Notes)
       - vital signs monitor (Results as per Appendix 2.2: Vital Signs)
       - cannulation trolley
       - IV fluids
       - printed ECG (Appendix 2.3: ECG #2- atrial fibrillation)
       - crash cart (standard set up)-out of view, brought in if required
       - medications available (aspirin, anginine, GTN patch, morphine, frusemide, metoprolol, clopidogrel)
       - oxygen- nasal prongs & Hudson mask
       - printed ABG/VBG results (Appendix 2.4: ABG/VBG results)

V. ACTORS
   A. Patient- 76 year old man/woman with dyspnoea
   B. Ward nurse
VI. CASE DETAILS

A. Initial information given to participant:
- You are a junior doctor on a medical cover shift at The Northern Hospital.
- You have just received the following page: “Pre-MET call, Unit A, Bed 8A.”
- You are in Unit A, so you go immediately to see the patient.

B. Handover from ward nurse
- “Mr(s) White only came up from ED a few hours ago. I was just doing his/her obs when (s)he told me that (s)he was feeling increasingly short of breath. His/her resp rate is up to 29 and his/her O2 sats had gone down to 90 on room air, so I’ve just put him/her on 2L of oxygen. I’m just doing the rest of his/her obs now.”

C. History from patient
1. Personal information
   - Andrew/Andrea White
   - 76 years old

2. HOPS
   * Volunteer the following information only:
     - “I’m just feeling very short of breath.”
   * Give the following information on direct questioning:
     * reason for being in hospital
       - “I’ve had this infection on my arm for a week or so. It wasn’t getting better so my GP sent me in here.”
     * dyspnoea
       - onset: had been feeling a little short of breath for a couple of hours but worsening
       - description: “hard to catch my breath”
     * associated symptoms
       - “feel like my heart is racing”
       - no chest pain
       - “dizzy”, if asked to explain: “lightheaded”
       - some cough, just since being in ED, productive of white sputum
       - sore right arm: “where the infection is”
   * systems review- all negative
     - exercise tolerance: walks to/from shops to do shopping (~20mins each way)
     - no orthopnoea/PND

3. Risk factors
   - HTN, T2DM (on OHAs)
   - unsure re: cholesterol
   - FHx: father died aged 85 of AMI; non-smoker, never smoked
   - for VTE
     - NO recent travel/surgery/immobility, NO personal Hx or FHx of VTE
     - for female pt: no HRT

4. PMHx
   - AF
   - HTN
   - T2DM
     - Dx 15 years ago
     - initially diet & exercise for a few years, then started on metformin
     - check BSLs mane & nocte, unsure of last HbA1c
   - OA: knees, hips
   - R TKR 2007
   - appendicectomy aged 19
5. Medications
   - perindopril 5mg daily
   - metoprolol 25mg BD
   - metformin 1g BD
   - panadol osteo ji TDS
   - warfarin (coumadin)- usual dose 5-6mg daily

6. Allergies
   - NKDA

7. Social Hx
   - home alone
   - widowed
   - 2 adult children nearby
   - retired accountant
   - never smoked, no household passive exposure
   - ETOH- “a sherry every night”

8. FHx - father AMI aged 85

D. Examination findings (given by assessor as participant performs that part of examination)
1. Appearance
   - appears dyspnoeic, increased work of breathing

2. Vital signs
   - on monitor- see Appendix 2.2

3. Neck- JVP at earlobe

4. Chest- apex beat not palpable, no heaves/thrills
   - HS dual, no added sounds
   - inspiratory crackles to midzones

5. Abdomen
   - SNT
   - appendicectomy scar RIF

6. Legs - pitting oedema to knees
   - calves SNT

7. Other- erythematous area on right arm- outline of cellulitis marked, erythema has not extended
   * No other positive examination findings

E. ECG
Appendix 2.3- ECG #2- Atrial Fibrillation with Rapid Ventricular Response
Appendix 5: Page 11 of 30

F. Nurse instructions:
   * ONLY perform tasks within your scope of practice
   * When performing tasks, do so in real time (ie do not perform the task faster than you really could)
   * Invasive tasks should be imitated only (eg. venepuncture, cannulation, administering medications)

   * Below is a list of responses to questions/requests from participants:
     1. Asked for vital signs
        - ask participant to refer to obs chart or monitor
     2. Asked about medications
        - direct participant to drug chart
     3. Asked to perform ECG
        - give participant ECG #2 initially
     4. Asked to take bloods
        - ask for pathology slip, ask if the patient needs a cannula
     5. Asked to insert a cannula
        - ask if the patient needs bloods taken too, ask for a pathology slip
     6. Asked to arrange a chest x-ray
        - ask for request form, imitate calling radiology to ask for mobile CXR
     7. Asked to give patient oxygen
        - ask how much; if Hudson mask not specified, use nasal prongs
     8. Asked to give any medications
        - ensure they are charted first
        - question any unusual doses/routes of administration
        - advise that an amiodarone infusion or IV metoprolol can only be given on telemetry
     9. Asked for previous records
        - hand participant the patient notes
     10. Asked to seek assistance of a senior doctor
        - return after 30 seconds, inform participant that you cannot find the registrar but you have paged them to attend
     11. Asked to call MET call
        - imitate calling MET call

F. Scenario progress

1. Dyspnoea
   - will improve *slightly* with oxygen or morphine
   - will not improve during simulation with oral frusemide
   - will improve with IV frusemide, but pt will remain dyspnoeic

2. Vital signs
   - see Appendix 2.2: Vital Signs

3. ECG
   - will remain in AF, rate unchanged

4. Examination findings
   - crackles will decrease, to lower zones only, with IV frusemide

5. Blood tests
   - results will not be available by the end of the simulation
   - if VBG is requested at time of venepuncture/cannulation, results will be provided in 3 minutes (Appendix 2.4: VBG)
   - if VBG requested after initial blood tests sent, repeat venepuncture will be required,
     results of VBG will take 3 minutes after venepuncture (Appendix 2.4: VBG)
   - if the participant performs an ABG, results will be available after 3 minutes (Appendix 1.4: ABG)

6. CXR
   - will not be performed by the end of the simulation

7. Request for assistance by registrar/consultant
   - assistance will arrive at 9 minutes

8. MET call
   - assistance will arrive 4 minutes after a MET call is called
Appendix 5: Page 12 of 30

VII. DEBRIEF
Five minutes total structured debrief

A. Structure
1. Ask the participant what they think they did well
2. Tell the participant which aspects they performed well on
3. Ask the participant what they think they could improve on
4. Tell the participant which aspects they performed less well on
5. Ask the participant what they thought was happening in the scenario
6. Explain basic management of acute myocardial infarction, if required

B. General guidelines
1. Do not divulge the participant’s scores
2. Do not compare the participant’s performance to that of other participants
3. If the participant appears to be distressed at their performance or the simulation outcome, please direct them to the student researcher (Dr Anastasia Vlachadis Castles).
APPENDICES

Appendix 1.1: Triage Notes

UR: 0123456
SMITH, John
DOB: 01/02/1953
Address: 123 Smith St, Epping, VIC, 3072
GP: Dr. Brown

TRIAGE NOTES: 62YO MAN PRESENTS W/CENTRAL CP 8/10 SEVERITY, NO RADIATION, ? ASSOC SOB.

OBSERVATIONS: DIAPHORETIC, HR 100 (reg), BP 160/110, RR 22, SpO2 95% RA, Temp 36.8

TRIAGE CATEGORY: 2
Appendix 1.1: Triage Notes

UR: 0123456
SMITH, Joan
DOB: 01/02/1953
Address: 123 Smith St, Epping, VIC, 3072
GP: Dr. Brown

TRIAGE NOTES: 62YO WOMAN PRESENTS W/CENTRAL CP 8/10 SEVERITY, NO RADIATION, ? ASSOC SOB.

OBSERVATIONS: DIAPHORETIC, HR 100 (reg), BP 160/110, RR 22, SpO2 95% RA, Temp 36.8

TRIAGE CATEGORY: 2
**Appendix 1.2: Vital Signs**

*Note: If asked for manual vital signs, results are same. There is no discrepancy in BP between arms.*

1. **Triage Vital Signs:**
   - HR 100 (reg)
   - BP 160/110
   - RR 22
   - SpO\textsubscript{2} 95% RA
   - Temp 36.8

2. **Initial Monitoring:**
   - HR 100 (reg)
   - BP 165/110
   - RR 26
   - SpO\textsubscript{2} 92% RA

3. **Changes with Treatment (5 minutes after treatment given):**
   - **A. Given CTN and Oxygen**
     - HR 90 (reg)
     - BP 130/90
     - RR 20
     - SpO\textsubscript{2} 95% if oxygen via NP
     - SpO\textsubscript{2} 97% if ≥ 8L/min oxygen via HM
   - **B. Given CTN but not oxygen**
     - HR 90 (reg)
     - BP 130/90
     - RR 24
     - SpO\textsubscript{2} 92% RA
   - **C. Given oxygen but not GTN**
     - HR 90 (reg)
     - BP 160/110
     - RR 20
     - SpO\textsubscript{2} 95% if oxygen via NP
     - SpO\textsubscript{2} 97% if oxygen via HM
   - **D. Not given CTN, nor oxygen**
     - HR 110 (reg)
     - BP 175/110
     - RR 28
   - **E. SpO\textsubscript{2} 90% RA**

4. **Further Changes in Vital Signs with Treatment:**
   - **A. Given further GTN**
     - Further S/L GTN drops BP by SBP 20, DBP 10
     - GTN patch (5mg) drops BP by SBP 30, DBP 15
     - GTN patch (10mg) drops BP by SBP 40, DBP 20
   - **B. Given further oxygen**
     - Increasing flow rate or changing from NP to HM will increase SpO\textsubscript{2} to 100%
   - **C. Given morphine**
     - Any reasonable dose of morphine will drop HR by 10, SBP by 20, DBP 10
   - **D. Given metoprolol**
     - Oral metoprolol will not have an effect on vital signs during the simulation
     - 5mg IV metoprolol will decrease HR by 20bpm and decrease BP by SBP 30, DBP 15
   - **E. Given frusemide**
     - Frusemide will not have an effect on vital signs during the simulation
Appendix 5: Page 16 of 30

http://www.learntheheart.com/assets/1/7/AnteriorSTEMI3.jpg
(Modified- removed extra rhythm strips at bottom of image)
### Appendix 1.4: ABG/VBG

#### Arterial Blood Gas

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Normal range</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.41</td>
<td>(7.35-7.45)</td>
</tr>
<tr>
<td>pCO₂</td>
<td>38</td>
<td>(35-48 mmHg)</td>
</tr>
<tr>
<td>pO₂</td>
<td>80</td>
<td>(83-108 mmHg)</td>
</tr>
<tr>
<td>HCO₃</td>
<td>25</td>
<td>(22-26 mmol/L)</td>
</tr>
<tr>
<td>Base Ex</td>
<td>0</td>
<td>(-2 - +3 mmol/L)</td>
</tr>
<tr>
<td>sO₂</td>
<td>95</td>
<td>(95.0-98.8 %)</td>
</tr>
<tr>
<td>Sodium</td>
<td>138</td>
<td>(135-145 mmol/L)</td>
</tr>
<tr>
<td>Potassium</td>
<td>4.2</td>
<td>(3.4-4.5 mmol/L)</td>
</tr>
<tr>
<td>Chloride</td>
<td>110</td>
<td>(98-114 mmol/L)</td>
</tr>
<tr>
<td>Anion gap</td>
<td>12</td>
<td>(10-20 mmol/L)</td>
</tr>
<tr>
<td>Ionised Ca</td>
<td>1.17</td>
<td>(1.15-1.27 mmol/L)</td>
</tr>
<tr>
<td>Glucose</td>
<td>8.3</td>
<td>(3.6-5.2 mmol/L)</td>
</tr>
<tr>
<td>L-lactate</td>
<td>1.1</td>
<td>(&lt;1.3 mmol/L)</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>143</td>
<td>(117-174 g/L)</td>
</tr>
<tr>
<td>Haematocrit</td>
<td>45</td>
<td>(35.0-51.0 %)</td>
</tr>
<tr>
<td>O₂Hb</td>
<td>95</td>
<td>(95.0-98.0 %)</td>
</tr>
<tr>
<td>CO₂Hb</td>
<td>0.5</td>
<td>(0.5-1.5 %)</td>
</tr>
<tr>
<td>MetHb</td>
<td>1.0</td>
<td>(&lt; 3.0 %)</td>
</tr>
<tr>
<td>HHb</td>
<td>2.0</td>
<td>(&lt; 5.1 %)</td>
</tr>
<tr>
<td>FiO₂</td>
<td>not specified</td>
<td>(%)</td>
</tr>
</tbody>
</table>

#### Venous Blood Gas

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Normal range</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.40</td>
<td>(7.35-7.45)</td>
</tr>
<tr>
<td>pCO₂</td>
<td>48</td>
<td>(35-48 mmHg)</td>
</tr>
<tr>
<td>pO₂</td>
<td>42</td>
<td>(83-108 mmHg)</td>
</tr>
<tr>
<td>HCO₃</td>
<td>25</td>
<td>(22-26 mmol/L)</td>
</tr>
<tr>
<td>Base Ex</td>
<td>0</td>
<td>(-2 - +3 mmol/L)</td>
</tr>
<tr>
<td>sO₂</td>
<td>72</td>
<td>(95.0-98.8 %)</td>
</tr>
<tr>
<td>Sodium</td>
<td>138</td>
<td>(135-145 mmol/L)</td>
</tr>
<tr>
<td>Potassium</td>
<td>4.2</td>
<td>(3.4-4.5 mmol/L)</td>
</tr>
<tr>
<td>Chloride</td>
<td>110</td>
<td>(98-114 mmol/L)</td>
</tr>
<tr>
<td>Anion gap</td>
<td>12</td>
<td>(10-20 mmol/L)</td>
</tr>
<tr>
<td>Ionised Ca</td>
<td>1.17</td>
<td>(1.15-1.27 mmol/L)</td>
</tr>
<tr>
<td>Glucose</td>
<td>8.3</td>
<td>(3.6-5.2 mmol/L)</td>
</tr>
<tr>
<td>L-lactate</td>
<td>1.8</td>
<td>(&lt;1.3 mmol/L)</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>143</td>
<td>(117-174 g/L)</td>
</tr>
<tr>
<td>Haematocrit</td>
<td>45</td>
<td>(35.0-51.0 %)</td>
</tr>
<tr>
<td>O₂Hb</td>
<td>75</td>
<td>(95.0-98.0 %)</td>
</tr>
<tr>
<td>CO₂Hb</td>
<td>0.5</td>
<td>(0.5-1.5 %)</td>
</tr>
<tr>
<td>MetHb</td>
<td>1.0</td>
<td>(&lt; 3.0 %)</td>
</tr>
<tr>
<td>HHb</td>
<td>2.0</td>
<td>(&lt; 5.1 %)</td>
</tr>
<tr>
<td>FiO₂</td>
<td>not specified</td>
<td>(%)</td>
</tr>
</tbody>
</table>
# Appendix 1.5 - Simulation Assessment Sheet

## ASSESSMENT SHEET

### SIMULATION SCENARIO 1: ACUTE MYOCARDIAL INFARCTION

<table>
<thead>
<tr>
<th>Participant Identification Number:</th>
<th>Assessor Identification Number:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Assessment Criteria:

#### A. Patient assessment

1. Asks about pain history
   - a) location: N / Y
   - b) radiation: N / Y
   - c) character: N / Y
   - d) severity: N / Y
   - e) onset/timing: N / Y

2. Asks about assoc. symptoms
   - a) dyspnoea: N / Y
   - b) nausea/vomiting: N / Y
   - c) leg pain: N / Y

3. Risk factors
   - a) asks about cardiovascular risk factors (HTN, dyslipidaemia, diabetes, smoking, PHx) 0-1 / 2-3 / 4-5

4. Other history
   - a) asks about PMHx IHD: N / Y

5. Examination
   - a) Vital signs: N / Y
   - b) JVP: N / Y
   - c) heart (auscultates): N / Y
   - d) lungs: N / Y
   - e) legs: N / Y

6. Investigation
   - a) requests ECG: N / Y
   - b) recognises STEMI: N / Y
   - c) requests blood tests (FBE, UeC, troponin minimum): N / Y
   - d) requests CXR: N / Y

#### B. Patient management

1. ACS management
   - a) ensures aspirin given: N / Y
   - b) gives GTN: N / Y
   - c) gives oxygen: N / Y
   - d) gives further GTN or morphine for ongoing pain: N / Y

2. Patient progress
   - a) re-evaluates pain severity: N / Y
   - b) re-evaluates vital signs: N / Y

3. Prioritisation/ recognition of limitations
   - a) seeks senior assistance: N / Y

### Score: ____ / 27
<table>
<thead>
<tr>
<th>Global Assessment Score (0-3):</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (adds nothing significant to assessment and management of patient)</td>
<td>□</td>
</tr>
<tr>
<td>1 (assesses history only, makes some effort to manage AMI)</td>
<td>□</td>
</tr>
<tr>
<td>2 (assesses the patient satisfactorily, institutes some management of AMI)</td>
<td>□</td>
</tr>
<tr>
<td>3 (optimal ED assessment and management of AMI)</td>
<td>□</td>
</tr>
</tbody>
</table>

Total Score: ____/ 30
Appendix 2.1: Patient Notes

Medical Admission (Smith-HMO2)

76yo man presents w/ R forearm cellulitis, not improved w/ oral antibiotics.

PMHx:
1. AF
   - CHADS = 3
   - on metoprolol & warfarin
   - known to TNH cardiology

2. T2DM
   - Dx 15yrs ago
   - initially diet & exercise- controlled
   - commenced on metformin after ~ 5yrs
   - managed by GP
   - ? last Hba1c
   - sees podiatrist, ophthalmologist
   - Cx- cataracts, ? diabetic neuropathy

3. HTN
   - on perindopril + metoprolol

4. OA
   - knees, hips
   - on panadol osteo
   - R TKR 2007, awaiting L TKR

5. R TKR 2007

6. Appendicectomy

Medications:
- perindopril 5mg daily
- metoprolol 25mg BD
- metformin 1g BD
- panadol osteo ii TDS
- warfarin (coumadin)- usual dose 3-6mg daily

Allergies:
- NKDA
Social Hx:
- Home alone
- Functionally independent, walks unaided
- Non-smoker
- ETOH - 1 unit/day

HOPC:
- Grazed R forearm on piece of wood while gardening ~ 2/52 ago
- Noticed increasing erythema/swelling around area ~ 1/52 ago
- Committed on oral anti by GP
- Re-presented to GP today with increasing pain in arm & febrile symptoms (feeling hot/cold, rigors)
- GP noticed worsening erythema, temp 38.5
  → Advised to attend ED

O/E:
- HR 100, BP 100/70, RR 15, SpO2 98% RA, temp 38.6
- Appears well, alert

Las:
- FBE - 138/15.6(11.2)/257
- UeC - 139/4.2/102/23
  - Cr 97/Ur 8.3/eGFR 84
- CRP - 192
- LFTs - NAD
- INR - 2.3

Imp:
- R forearm cellulitis surrounding area of injury; systemically unwell
Plan:
1. Admit Med 4
2. Commence IV flucloxacillin
3. Daily dressings to grazed area
4. Mark area of erythema
5. Regular meds
6. IV fluids

Nursing Notes
Pt transferred from ED. Obs stable, febrile 38.6 in ED. IVC in situ L cubital fossa (inserted in ED). IV fluids running. Dx - R forearm cellulitis. Commenced on IV antis. Pt c/o pain to right forearm 6/10 on arrival to ward, given oxynorm with good effect. A. Johnson - RN1
Appendix 5: Page 23 of 30

Appendix 2.1: Patient Notes

**Medical Admission (Smith- HMO2)**

76yo woman presents w/ R forearm cellulitis, not improved w/ oral antibiotics.

**PMHx**
1. AF
   - CHADS = 3
   - on metoprolol & warfarin
   - known to TNH cardiology

2. T2DM
   - Dx 15yrs ago
   - initially diet & exercise- controlled
   - commenced on metformin after ~ 5yrs
   - managed by GP
   - ? last Hbr10c
   - sees podiatrist, ophthalmologist
   - Cx- cataracts, ? diabetic neuropathy

3. HTN
   - on perindopril + metoprolol

4. OA
   - knees, hips
   - on panadol osteo
   - R TKR 2007, awaiting L TKR

5. R TKR 2007

6. Appendicectomy

**Medications:**
- perindopril 5mg daily
- metoprolol 25mg BD
- metformin 1g BD
- panadol osteo ii TDS
- warfarin (coumadin) - usual dose 5-6mg daily

**Allergies:**
- NXDA
Social Hx:
Home alone
Functionally independent, walks unaided
Non-smoker
ETOH - 1 unit/day

HOPS:
- Grazed R forearm on piece of wood while gardening ~ 2/52 ago
- Noticed increasing erythema/swelling around area ~ 1/52 ago
- Commenced on oral antibiotics by GP
- Re-presented to GP today with increasing pain in arm & febrile symptoms
  (feeling hot/cold, rigors)
- GP noticed worsening erythema, temp 38.5
  Advised to attend ED

O/E:
HR 100, BP 100/70, RR 15, SpO2 98% RA, temp 38.6
Appears well, alert

Ix:
FBE - 138/15.6(11.2)/257
UeC - 139/4.2/102/23
  Cr 97/Ur 8.3/eGFR 84
CRP - 192
LFTs - NAD
INR - 2.3

Imp:
R Forearm cellulitis surrounding area of injury; systemically unwell
Plan:
1. Admit Med 4
2. Commence IV flucloxacillin
3. Daily dressings to grazed area
4. Mark area of erythema
5. Regular meds
6. IV fluids

Nursing Notes
Pt transferred from ED. Obs stable, febrile 38.6 in ED. IVC in situ L cubital fossa (inserted in ED). IV fluids running. Dx - R forearm cellulitis. Commenced on IV anti. Pt c/o pain to right forearm 6/10 on arrival to ward, given oxynorm with good effect. A. Johnson – RN1
Appendix 2.2: Vital Signs

*Note: If asked for manual vital signs, results are same. There is no discrepancy in BP between arms.

1. **Previous Observations:**
   See patient notes

2. **Initial Monitoring:**
   - HR 130 (irreg)
   - BP 110/70
   - RR 29
   - SpO2 92% on 2L via NP
   - temp 37.8

3. **Changes with Treatment**
   A. Given further oxygen
      Increasing flow rate or changing from NP to HM (at ≥ 8L/min) will decrease RR to 26, SpO2 remains 92% on O2
   B. Given frusemide
      Oral frusemide will not have an effect on vital signs during the simulation
      IV frusemide will not have an effect on HR/BP during the simulation
      IV frusemide 20mg, after 5 mins → RR 26, SpO2 92% on O2 (90% on RA)
      IV frusemide 40mg, after 5 mins → RR 24, SpO2 92% on O2 (90% on RA)
   C. Given metoprolol
      Metoprolol will not have an effect on vital signs during the simulation
   D. Given morphine
      HR 130 (irreg)
      BP 100/650
      RR 24
      SpO2 92% on O2 (90% on RA)
      * if also given frusemide:
      IV frusemide 20mg, after 5 mins → RR 26, SpO2 92% on O2 (90% on RA)
      IV frusemide 40mg, after 5 mins → RR 24, SpO2 92% on O2 (90% on RA)
   E. Given GTN
      HR 130 (irreg)
      BP 90/50
      RR 24
      SpO2 92% on O2 (90% on RA)
      * if also given frusemide:
      IV frusemide 20mg, after 5 mins → RR 26, SpO2 92% on O2 (90% on RA)
      IV frusemide 40mg, after 5 mins → RR 24, SpO2 92% on O2 (90% on RA)
Appendix 2.3: ECG #2 - Atrial Fibrillation

ECG #2 - Atrial Fibrillation with Rapid Ventricular Rate

From http://lifeinthefastlane.com/ecg-library/atrial-fibrillation/
### Appendix 2.4: ABG/VBG Results

#### Arterial Blood Gas

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Normal Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.41</td>
<td>(7.35-7.45)</td>
</tr>
<tr>
<td>pCO2</td>
<td>38</td>
<td>(35-48 mmHg)</td>
</tr>
<tr>
<td>pO2</td>
<td>75</td>
<td>(83-108 mmHg)</td>
</tr>
<tr>
<td>HCO3</td>
<td>25</td>
<td>(22-28 mmol/L)</td>
</tr>
<tr>
<td>Base Ex</td>
<td>0</td>
<td>(-2 - +3 mmol/L)</td>
</tr>
<tr>
<td>sO2</td>
<td>90</td>
<td>(95.0-98.8 %)</td>
</tr>
<tr>
<td>Sodium</td>
<td>138</td>
<td>(135-145 mmol/L)</td>
</tr>
<tr>
<td>Potassium</td>
<td>4.2</td>
<td>(3.4-4.5 mmol/L)</td>
</tr>
<tr>
<td>Chloride</td>
<td>105</td>
<td>(98-114 mmol/L)</td>
</tr>
<tr>
<td>Anion gap</td>
<td>12</td>
<td>(10-20 mmol/L)</td>
</tr>
<tr>
<td>Ionised Ca</td>
<td>1.17</td>
<td>(1.15-1.27 mmol/L)</td>
</tr>
<tr>
<td>Glucose</td>
<td>8.3</td>
<td>(3.6-5.2 mmol/L)</td>
</tr>
<tr>
<td>L-lactate</td>
<td>1.1</td>
<td>(&lt;1.3 mmol/L)</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>140</td>
<td>(117-174 g/L)</td>
</tr>
<tr>
<td>Haematocrit</td>
<td>45</td>
<td>(35.0-51.0 %)</td>
</tr>
<tr>
<td>O2Hb</td>
<td>90</td>
<td>(95.0-98.0 %)</td>
</tr>
<tr>
<td>CO2Hb</td>
<td>0.5</td>
<td>(0.5-1.5 %)</td>
</tr>
<tr>
<td>MetHb</td>
<td>1.0</td>
<td>(&lt; 3.0 %)</td>
</tr>
<tr>
<td>HHb</td>
<td>2.0</td>
<td>(&lt; 5.1 %)</td>
</tr>
<tr>
<td>FiO2</td>
<td>not specified</td>
<td>(%)</td>
</tr>
</tbody>
</table>

#### Venous Blood Gas

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Normal Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.40</td>
<td>(7.35-7.45)</td>
</tr>
<tr>
<td>pCO2</td>
<td>48</td>
<td>(35-48 mmHg)</td>
</tr>
<tr>
<td>pO2</td>
<td>38</td>
<td>(83-108 mmHg)</td>
</tr>
<tr>
<td>HCO3</td>
<td>25</td>
<td>(22-28 mmol/L)</td>
</tr>
<tr>
<td>Base Ex</td>
<td>0</td>
<td>(-2 - +3 mmol/L)</td>
</tr>
<tr>
<td>sO2</td>
<td>70</td>
<td>(95.0-98.8 %)</td>
</tr>
<tr>
<td>Sodium</td>
<td>138</td>
<td>(135-145 mmol/L)</td>
</tr>
<tr>
<td>Potassium</td>
<td>4.2</td>
<td>(3.4-4.5 mmol/L)</td>
</tr>
<tr>
<td>Chloride</td>
<td>105</td>
<td>(98-114 mmol/L)</td>
</tr>
<tr>
<td>Anion gap</td>
<td>12</td>
<td>(10-20 mmol/L)</td>
</tr>
<tr>
<td>Ionised Ca</td>
<td>1.17</td>
<td>(1.15-1.27 mmol/L)</td>
</tr>
<tr>
<td>Glucose</td>
<td>8.3</td>
<td>(3.6-5.2 mmol/L)</td>
</tr>
<tr>
<td>L-lactate</td>
<td>1.8</td>
<td>(&lt;1.3 mmol/L)</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>140</td>
<td>(117-174 g/L)</td>
</tr>
<tr>
<td>Haematocrit</td>
<td>45</td>
<td>(35.0-51.0 %)</td>
</tr>
<tr>
<td>O2Hb</td>
<td>70</td>
<td>(95.0-98.0 %)</td>
</tr>
<tr>
<td>CO2Hb</td>
<td>0.5</td>
<td>(0.5-1.5 %)</td>
</tr>
<tr>
<td>MetHb</td>
<td>1.0</td>
<td>(&lt; 3.0 %)</td>
</tr>
<tr>
<td>HHb</td>
<td>2.0</td>
<td>(&lt; 5.1 %)</td>
</tr>
<tr>
<td>FiO2</td>
<td>not specified</td>
<td>(%)</td>
</tr>
</tbody>
</table>
### Appendix 2.5- Simulation Assessment Sheet

**ASSESSMENT SHEET SIMULATION SCENARIO 2: ACUTE PULMONARY OEDEMA**

<table>
<thead>
<tr>
<th>Assessment Criteria:</th>
<th>Participant Name:</th>
</tr>
</thead>
</table>

| Participant Identification Number: | ____________________________ | 
| Assessor Identification Number: | ____________________________ | 

#### A. Patient assessment

1. Asks about dyspnoea
   - a) onset
   - b) exertional/at rest

2. Asks about assoc. symptoms
   - a) palpitations
   - b) chest pain
   - c) lightheadedness
   - d) cough
   - e) sputum
   - f) orthopnoea
   - g) PND

3. Examination
   - a) Vital signs
   - b) JVP
   - c) heart sounds
   - d) lungs
   - e) legs

4. Investigation
   - a) requests ECG
   - b) recognises AF
   - c) requests CXR
   - d) requests bloods
     (FBE, U/E as minimum)
   - e) requests ABG/VBG

#### B. Patient management

1. ACS management
   - a) sits patient up
   - b) gives oxygen
   - c) gives IV frusemide
   - d) gives GTN and/or morphine and/or metoprolol
   - e) stops IV fluids

2. Patient progress
   - a) re-evaluates dyspnoea
   - b) re-evaluates vital signs

3. Prioritisation/recognition of limitations
   - a) seeks senior assistance

**Score:** __/27
Global Assessment Score (0-3):

- 0 (adds nothing significant to assessment and management of patient)
- 1 (assesses history only, makes some effort to manage fluid overload)
- 2 (assesses the patient satisfactorily, institutes some management of fluid overload)
- 3 (optimal initial ward assessment and management of APO)

Total Score: ___/30
Appendix 6: Recruitment Material

Are you interested in being involved in educational research looking at SIMULATION as a method for assessing MANAGEMENT OF CARDIAC EMERGENCIES?

A doctor based at The Northern Hospital is conducting research into using simulation to assess management of cardiac emergencies.

Medical students and JMOs are being sought to participate.
Participation is completely voluntary and will involve:
- individual simulation experience, including feedback
- completion of a thirty minute MCQ test on the same day
- completion of a 5 minute participant information questionnaire
→ total time commitment ~ 1hr (all on the same day)

This is a great opportunity to receive simulation-based experience in crucial cardiac emergency management skills. You would also be assisting in educational research, designed to help bring innovative teaching and assessment techniques into practice.

If you think you might be interested, please write your name & email address on the sheet below, and you will be provided with further information. If you have any questions please contact Dr. Anastasia Castles at anastasia.castles@research.uwa.edu.au
If you think you might be interested, please write your name & best contact email below, and you will be provided with further information via email.

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
<tr>
<td>_____________</td>
<td>______________</td>
</tr>
</tbody>
</table>
Sample Email:

Dear student/JMO,

This is an exciting educational opportunity for you!
Are you interested in being involved in educational research looking at SIMULATION as a method for assessing MANAGEMENT OF CARDIAC EMERGENCIES? 
A doctor based at The Northern Hospital is conducting research into using simulation to assess management of cardiac emergencies.

Medical students and JMOs are being sought to participate.

Participation is completely voluntary and will involve:
- individual simulation experience, including feedback
- completion of a thirty minute MCQ test on the same day
- completion of a 5 minute participant information questionnaire

→ total time commitment ~ 1hr (all on the same day)

This is a great opportunity to receive simulation-based training in crucial cardiac emergency management skills. You would also be assisting in educational research, designed to help bring innovative teaching and assessment techniques into practice.

If you think you might be interested, please send a quick email with your name, year level and best contact email address to Dr. Anastasia Castles:
anastasia.castles@research.uwa.edu.au

You will then be provided with further information. Your participation in the simulation training session can be timed to fit around your timetable/roster. Any questions can also be directed to the above email address.

Kind regards,
Dr. Anastasia Castles