Assessing decision self-monitoring using item response certainty and safeness

A thesis submitted for the degree of
Doctor of Philosophy
University of Otago

Michael Tweed, December 2021
Abstract

Diagnostic and therapeutic decision-making is perhaps the most important role of doctors as healthcare professionals. The majority of these judgements by doctors are correct, but when erroneous decisions are made it can lead to adverse events with significant implications for patient welfare. Doctors require not only sufficient knowledge and skill to make decisions, but also the capacity to know when they have reached the limit of that knowledge and skill. This capacity is self-monitoring, which is the activity of reflection-in-action as part of decision-making in daily practice. As it is a critical component of practice, medical education needs to assist students to develop self-monitoring.

One approach to developing self-monitoring is to embed it into medical education assessment. The main purposes of assessment include guiding and motivating learning and informing decisions on whether students should progress to subsequent levels. Research into assessment related to self-monitoring is relatively sparse. Belief about one’s ability to undertake actions and about the consequences of those actions are at the core of the theories that underpin self-monitoring. Assessment of self-monitoring could be informed by evaluating certainty in assessment item responses and safety of item responses.

There is a need to develop an assessment technique for self-monitoring that would include items based on authentic clinical practice and that would incorporate both certainty in, and safety of, item responses. Certainty could be assessed by having students rate their certainty in their multiple-choice examination responses. Safety could be assessed by having experts rate the potential outcome of each option in a multiple-choice item.

This thesis includes four separate studies. Across these studies medical students sat a variety of multiple-choice tests which were of different stakes, length and content. They were designed to introduce and evaluate the inclusion of certainty in, and safety of, responses. The studies included a variety of outcomes with evaluation at cohort and individual student level for: feasibility; scoring systems; progression decision-making; correctness for levels of certainty; safeness for levels of certainty; and the influence of student ability, experience, gender, ethnicity, and feedback on correctness-in and safeness-of responses for levels of certainty.
The main findings were that:

1) It was feasible to incorporate certainty in responses, and to a lesser extent, safeness of responses.

2) The reliability of responses was greatest for certainty, less for correctness and least for safeness.

3) Students of all levels of ability and experience demonstrated increasing correctness with increasing certainty.

4) Unsafe responses occurred across all levels of ability, experience and certainty.

Correctness, correctness for levels of certainty, and safeness for levels of certainty have been demonstrated to vary in different ways, suggesting that they measure different constructs and therefore could be important to introduce to medical student education programmes. Multiple-choice question tests with certainty-in and safeness-of responses introduced early in medical education and training is a potential means to better understand and enhance learning of self-monitoring of clinical decisions.
Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be at the front (after the thesis abstract) of the thesis submitted for examination and library deposit.

Expectations of theses including publications:

The thesis must be an integrated and coherent body of work. The body of the thesis may include entire publications verbatim, but these need to be reformatted to a consistent chapter style, and commentary may need to be added to link the publications together. If the thesis contains multiple publications, candidates need to ensure that they have: included an introductory chapter; demonstrated their ability to critically engage with the literature; carefully considered their research design and justified choices of methodology and methods where necessary; and synthesised and discussed the findings from the publications. It is the candidate’s responsibility to ensure that any published work (or parts thereof) included in the thesis comply with the copyright provisions of the publisher and that any guidelines with regard to self-citation are followed.

Details of publications included in and/or appended to this thesis (please add rows as needed).

<table>
<thead>
<tr>
<th>Chapter/Append.</th>
<th>Paper title</th>
<th>Authors</th>
<th>Contribution of candidate and co-authors – please detail the nature and extent (%)</th>
<th>Journal</th>
<th>Status (e.g. under review, forthcoming, published)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017, 51(3), 316-323</td>
<td>Low performing students have insightfulness when they reflect-in-action.</td>
<td>Tweed, M., Purdie, G., &amp; Wilkinson, T.</td>
<td>MT (candidate) conceived the idea, the original study design, developed the assessments, and made first interpretation of results. GP undertook the majority of the statistical analysis. TW advised throughout the process. MT made first draft for submission and first re-draft for resubmission following reviewers comments. MT (85%), GP (5%), TW (10%)</td>
<td>Medical Education</td>
<td>Published</td>
</tr>
<tr>
<td>Chapter/Append.</td>
<td>Paper title</td>
<td>Authors</td>
<td>Contribution of candidate and co-authors – please detail the nature and extent (%)</td>
<td>Journal</td>
<td>Status (e.g. under review, forthcoming, published)</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>---------------------------------------------------</td>
</tr>
<tr>
<td>2017, 17(1), 1-11.</td>
<td>Certainty and safe consequence responses provide additional information from multiple choice question assessments.</td>
<td>Tweed, M. J., Stein, S., Wilkinson, T. J., Purdie, G., &amp; Smith, J</td>
<td>MT (candidate) conceived the idea, the original study design, developed the assessments, and made first interpretation of results. GP undertook the majority of the statistical analysis. SS, TJW, JS advised throughout the process. MT made first draft for submission and first re-draft for resubmission following reviewers comments.</td>
<td>BMC Medical Education</td>
<td>Published</td>
</tr>
<tr>
<td>2013, 35(2), 127-133.</td>
<td>Determining measures of insight and foresight from responses to multiple choice questions.</td>
<td>Tweed, M., Thompson-Fawcett, M., Schwartz, P., &amp; Wilkinson, T. J.</td>
<td>MT (candidate) conceived the idea, the original study design, developed the assessments, and made first interpretation of results. MTF, PS, TJW advised throughout the process. MT made first draft for submission and first re-draft for resubmission following reviewers comments.</td>
<td>Medical Teacher</td>
<td>Published</td>
</tr>
<tr>
<td>Chapter/Append.</td>
<td>Paper title</td>
<td>Authors</td>
<td>Contribution of candidate and co-authors – please detail the nature and extent (%)</td>
<td>Journal</td>
<td>Status (e.g. under review, forthcoming, published)</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>2012, 13(3), 84-92</td>
<td>A confidence and safety approach to multiple-choice question scoring.</td>
<td>Tweed, M. J., Thompson-Fawcett, M., Schwartz, P., &amp; Wilkinson, T. J.</td>
<td>MT (candidate) conceived the idea, the original study design, developed the assessments, and made first interpretation of results. MTF, PS, TJW advised throughout the process. MT made first draft for submission and first re-draft for resubmission following reviewers comments. MT (80%), MTF (5%), PS (5%), TJW (10%)</td>
<td>Focus on Health Professional Education: A Multi-disciplinary Journal</td>
<td>Published</td>
</tr>
</tbody>
</table>

**Certification by Primary Supervisor:**

The undersigned certifies that the above table correctly reflects the nature and extent of the candidate’s contribution to this co-authored work.

Name: Jeffrey Smith  
Signature:  
Date: 12/1/21
Acknowledgements

I would like to acknowledge the many people who have supported me on my journey.

Firstly, my supervisors Professor Jeff Smith, Professor Tim Wilkinson, and Dr Sarah Stein, for their guidance and assistance.

I would like to acknowledge technical support received from Mr Gordon Purdie and Dr James Stanley for assistance in statistical analysis, from Mr Adon Moskal for software development for examination delivery, and from Ms Fiona Hyland for support proof reading and editing drafts of this thesis.

I am grateful to the students and staff for their active involvement in developing, undertaking and providing comment on various aspects of the project. Likewise the administration staff who organised staff and students.

I am grateful to the Department of Medicine, University of Otago Wellington, the College of Education, University of Otago, and the Otago Medical School who provided financial support to assist with some of the expenses associated with this thesis.

Thanks too, to my local, national and international colleagues who, as conference audiences, journal editors and reviewers and friends, provided valuable comment and maintained my enthusiasm.

Lastly to my sons Jon and Nick and especially my partner Karin: thank you for ongoing support throughout this time.
Table of Contents

Abstract .................................................................................................................................................. i
Acknowledgements............................................................................................................................. vi
List of Tables ......................................................................................................................................... xvii
List of Figures ......................................................................................................................................... xxi

Chapter 1. Preface and overview ............................................................................................................ 22
  1.1 Preface............................................................................................................................................. 23
  1.2 Overview ......................................................................................................................................... 24

Chapter 2. Introduction to self-monitoring and the concept of self-efficacy ............................................. 26
  2.1 Chapter outline................................................................................................................................. 27
  2.2 Self-monitoring............................................................................................................................... 28
  2.3 The concept of self-efficacy ............................................................................................................. 29
    2.3.1 Self-efficacy within social cognitive theories................................................................. 29
    2.3.2 Self-efficacy theory ............................................................................................................... 30
    2.3.3 Self-efficacy in healthcare professional practice ............................................................ 32
    2.3.4 Self-efficacy and medical education ..................................................................................... 34
    2.3.5 Summary ............................................................................................................................... 35
  2.4 Self-monitoring for healthcare professions ................................................................................... 36
    2.4.1 A research agenda .................................................................................................................. 36
    2.4.2 Self-monitoring in clinical practice ....................................................................................... 36
    2.4.3 Integrating self-monitoring into clinical practice, training and education ...................... 40
    2.4.4 Learning and developing self-monitoring for clinical practice 40
    2.4.5 Self-monitoring in medical degree curricula ....................................................................... 45
    2.4.6 Summary ............................................................................................................................... 45
  2.5 Healthcare professionals’ clinical decision-making ........................................................................ 46
    2.5.1 Erroneous decisions by individual clinicians ....................................................................... 46
    2.5.2 Individual ability and error .................................................................................................... 47
    2.5.3 Institutional, system and individual factors relating to patient safety ............................... 48
    2.5.4 Individual approaches ............................................................................................................ 49
    2.5.5 Summary ............................................................................................................................... 49
2.6 Validation of self-monitoring as an important part of clinical practice

.................................................................50

2.6.1 Environmental scan.................................................................50

2.6.2 Summary .................................................................................52

2.7 Summary for Chapter ..................................................................53

Chapter 3.  Assessment with meta-judgements, item level certainty, item response safeness, and self-monitoring..........................................................54

3.1 Chapter outline..............................................................................55

3.2 Introduction to assessment..............................................................56

3.2.1 Purposes of assessment.............................................................56

3.3 Assessment of self-efficacy .............................................................58

3.3.1 Self-efficacy measurement: comparison of perceived capability and actual performance.................................................................58

3.3.2 Self-efficacy measurement: Generality/Specificity .......................59

3.3.3 Self-efficacy measurement: consequences ...................................60

3.3.4 Self-efficacy measurement: assessment providing educational value.........................................................................................60

3.3.5 Self-efficacy measurement: aggregation........................................61

3.4 Assessment of meta-judgements of cognitive capability ...................63

3.4.1 Prediction of global performance made before performance ..........63

3.4.2 Post-diction of global performance.............................................64

3.4.3 Post-diction of item level performance made after performance: calibration...................................................................................64

3.4.4 Miscalibration ............................................................................65

3.4.5 Calibration and education ...........................................................66

3.4.6 Limitation of some calibration analyses for HCP practice..............67

3.4.7 Summary of assessment of self-monitoring as a meta-judgements of cognitive capability .........................................................68

3.5 Item-level certainty ........................................................................69

3.5.1 Which words? ............................................................................69

3.5.2 Origins of CBA..........................................................................69

3.5.3 Developments of CBA.................................................................70

3.5.4 Reported advantages and potential problems of CBA...............71

3.5.5 Methods to measure certainty ....................................................73

3.5.6 Methods of scoring based on certainty ........................................82
3.5.7 Applications of CBA scores ...........................................86
3.5.8 Certainty, gambling, and risk-taking ..............................89
3.5.9 Item level certainty and HCP clinical decision-making ...........91
3.5.10 Evaluation of CBA by students ......................................91
3.5.11 Content specificity, judgements and meta-judgements ..........92
3.5.12 Non MCQ tests ..........................................................92
3.5.13 Summary ...................................................................93

3.6 Safeness of responses .......................................................94
3.6.1 Unsafe answers in written assessments ............................94
3.6.2 Aggregating unsafe answers into assessment decisions ........98
3.6.3 High-stakes decisions based on a low number of unsafe
       responses ........................................................................100
3.6.4 Tolerance for error as a learning opportunity ....................101
3.6.5 Certainty in unsafe responses being correct .....................101
3.6.6 Summary ...................................................................102

3.7 Specific assessment related to self-monitoring ...................103
3.7.1 Evidence for self-monitoring inferred from assessment
       responses ........................................................................103
3.7.2 Assessment of individuals for evidence for their self-
       monitoring .....................................................................108
3.7.3 Summary of assessment of self-monitoring ....................125

3.8 Summary of the chapter ..................................................126

Chapter 4. Research paradigm and context..........................127
4.1 Chapter outline ...............................................................128
4.2 Research paradigm ..........................................................129
4.3 Ethics committee approvals and research consultation with Māori .131
4.4 Otago Medical School (OMS) MBChB degree Course (relevant years
       2008-2015) ..................................................................132
4.4.1 Course overview ..........................................................132
4.4.2 ELM purpose ...............................................................132
4.4.3 ELM overview .............................................................133
4.4.4 ALM purpose ...............................................................134
4.4.5 ALM overview .............................................................135
4.4.6 Whole class assessments ..............................................136
4.5 Summary for chapter ......................................................139
Chapter 5. Developing assessment for certainty-in and safeness-of responses ..140

5.1 Chapter outline ......................................................................................................................................141
5.2 Formats of assessment .................................................................................................................................142
5.3 MCQs .............................................................................................................................................................143
  5.3.1 MCQ stems/topics .......................................................................................................................................143
  5.3.2 MCQ lead-ins .............................................................................................................................................143
  5.3.3 MCQ response options ...............................................................................................................................144
  5.3.4 Extended matching MCQ .........................................................................................................................145
  5.3.5 Long menu options ..................................................................................................................................146
  5.3.6 Order of presentation of stem, certainty judgement and options ................................................................146
  5.3.7 MCQ scoring .............................................................................................................................................147
  5.3.8 ‘Don’t know’ option ..................................................................................................................................149
  5.3.9 Guessing and scoring ...............................................................................................................................149
  5.3.10 Strategic responding, test-wiseness and personality ...........................................................................150
  5.3.11 Making decisions in clinical practice and making decisions to respond to MCQs..............................152
5.4 Writing quality MCQs ..................................................................................................................................153
5.5 Standard setting ..............................................................................................................................................154
  5.5.1 Standard setting for knowledge ...............................................................................................................154
  5.5.2 Standard setting certainty, safeness, interactions with correctness and self-monitoring .................155
5.6 Feedback to guide learning ............................................................................................................................156
5.7 Summary for chapter ....................................................................................................................................159

Chapter 6. Study 1 ..............................................................................................................................................160

6.1 Introduction ...................................................................................................................................................161
  6.1.1 Hypotheses and research questions ......................................................................................................161
6.2 Methods ........................................................................................................................................................163
  6.2.1 Design ......................................................................................................................................................163
  6.2.2 Question pool ..........................................................................................................................................163
  6.2.3 Paper development ................................................................................................................................165
  6.2.4 Confidence (certainty) descriptors .........................................................................................................166
  6.2.5 Subjects: candidates, recruitment, consequence/stakes ....................................................................168
  6.2.6 Delivery format ......................................................................................................................................168
  6.2.7 Information for candidates ...................................................................................................................168
7.2.6 Subjects: candidates, recruitment, consequence/stakes........202
7.2.7 Delivery format.........................................................202
7.2.8 Information for students.............................................204
7.2.9 Feedback to students..................................................204
7.3 Analysis............................................................................205
7.3.1 Response and scoring analysis........................................205
7.3.2 Analysis linked to research questions..............................205
7.4 Results..............................................................................207
7.4.1 General response data....................................................207
7.4.2 Relationship between levels of certainty and the odds of correctness.................................................210
7.4.3 Relationship between levels of certainty and the odds of correctness by ability group..........................210
7.4.4 Relationship between levels of certainty and the odds of unsafeness among all (correct and incorrect) responses ........211
7.4.5 Relationship between levels of certainty and the odds of unsafeness among all (correct and incorrect) responses by ability group..................................................211
7.4.6 Relationship between levels of certainty and the odds of an unsafe response among incorrect responses...............212
7.4.7 Relationship between levels of certainty and the odds of an unsafe response among incorrect responses by ability group213
7.4.8 Relationship between levels of certainty and the odds of moderate or highly unsafe responses among all (correct and incorrect) responses.................................................214
7.4.9 Relationship between levels of certainty the odds of moderate or highly unsafe responses among all (correct and incorrect) responses by score group.............................................215
7.4.10 Relationship between levels of certainty the odds of moderate or highly unsafe responses among incorrect responses ........215
7.4.11 Relationship between levels of certainty the odds of moderate or highly unsafe responses among incorrect responses by ability group..................................................216
7.5 Discussion........................................................................217
7.5.1 Feasibility of certainty-in and safeness-of responses..........218
7.5.2 Correctness for levels of certainty ........................................ 219
7.5.3 Safeness for levels of certainty ........................................... 220
7.5.4 Findings not related to research questions .......................... 221
7.5.5 Strengths of this Study .................................................. 221
7.5.6 Limitations of this study ................................................ 221

Chapter 8. Study 3 .................................................................... 223
8.1 Introduction .......................................................................... 224
  8.1.1 Building on prior studies ............................................. 224
  8.1.2 Hypotheses and research questions ............................. 225
8.2 Methods ............................................................................. 227
  8.2.1 Design ........................................................................ 227
  8.2.2 Question pool .............................................................. 227
  8.2.3 Question classification for safeness process .................. 228
  8.2.4 Question selection for use in the examinations ............. 229
  8.2.5 Subjects: candidates, recruitment, consequence/stakes .... 230
  8.2.6 Test format and provision of information to candidates ..... 230
  8.2.7 Feedback to students .................................................. 232
8.3 Analysis of test responses ................................................... 233
  8.3.1 Response and scoring analysis ...................................... 233
  8.3.2 Comparison between and among groups ....................... 233
  8.3.3 Correctness and safeness by certainty, for ability groupings 233
  8.3.4 Results by gender and ethnicity .................................... 234
  8.3.5 Associations between study (certainty and safety) test results and concurrent OMS assessment results 235
  8.3.6 Risk taking ................................................................. 235
  8.3.7 Effect of time/feedback ................................................ 236
8.4 Student interview process ..................................................... 237
8.5 Results ................................................................................. 238
  8.5.1 Correctness for levels of certainty ................................. 238
  8.5.2 Unsafeness for levels of certainty ................................... 239
  8.5.3 Effect of ability on responses, Test 1 ............................. 241
  8.5.4 Effect of student ethnicity on responses, Test 1 ................ 241
  8.5.5 Effect of student gender on responses, Test 1 ............... 242
  8.5.6 Comparisons with standard OMS assessments ............. 246
  8.5.7 Risk-taking ................................................................. 256
8.5.8  Change in responses between Test 1 and Test 2 ..................264
8.5.9  Student interview results ..................................................269
8.5.10 Interviews results .........................................................270
8.5.11 Themes .................................................................270

8.6  Discussion ........................................................................279
8.6.1 Feasibility of certainty-in and safeness-of responses .............281
8.6.2 Correctness for levels of certainty ........................................282
8.6.3 Safeness for levels of certainty ............................................282
8.6.4 Lack of effect of feedback on proportions correct or unsafe for given levels of certainty ........................................283
8.6.5 Risk-taking ..................................................................284
8.6.6 Correlations other assessments, gender and ethnicity ............286
8.6.7 Further interpretations of themes from interviews ..................287
8.6.8 Findings not related to the research questions .......................292
8.6.9 Strengths of this study .....................................................293
8.6.10 Limitations of this study ..................................................293

Chapter 9. Study 4 ..................................................................296
9.1  Introduction ........................................................................297
9.1.1 Building on Studies 1-3 .......................................................297
9.1.2 Developments addressed in Study 4 .....................................297
9.1.3 Hypotheses and research questions .....................................298

9.2  Methods ..............................................................................300
9.2.1 Design ...........................................................................300
9.2.2 Question pool .................................................................300
9.2.3 Question selection for use in the tests .................................301
9.2.4 Subjects: candidates, recruitment, consequences/stakes ........301
9.2.5 Delivery format ...............................................................302
9.2.6 Information for students ....................................................304
9.2.7 Feedback to students ........................................................304

9.3  Analysis ..............................................................................305
9.3.1 Response and scoring reliability .........................................305
9.3.2 Analysis linked to research questions ..................................305

9.4  Results ...............................................................................306
9.4.1 General data ..................................................................306
9.4.2 Results related to research questions .................................307
9.5 Discussion ............................................................................................................. 311
  9.5.1 Hypotheses ..................................................................................................... 311
  9.5.2 Research questions ....................................................................................... 311
  9.5.3 Feasibility of certainty-in and safeness-of responses ...................................... 312
  9.5.4 Correctness for given levels of certainty ....................................................... 313
  9.5.5 Findings not related to the research questions ............................................. 314
  9.5.6 Strengths of this study ................................................................................... 315
  9.5.7 Limitations of this study ............................................................................... 315

Chapter 10. General discussion and conclusions .................................................... 316
  10.1 Chapter outline ................................................................................................ 317
  10.2 Summary of findings ...................................................................................... 318
  10.3 Analysis of findings in relationship to prior research literature ...................... 320
    10.3.1 Assessment incorporating response certainty, and to a lesser extent incorrect response safety, was feasible ................................. 320
    10.3.2 Reliability of responses is greatest for certainty, less for correctness and least for safeness ................................................................. 320
    10.3.3 Students at all levels of ability demonstrate increasing correctness with increasing certainty .......................................................... 323
    10.3.4 Unsafe responses occurred across all level of student ability ................... 325
  10.4 Potential limitations ......................................................................................... 328
    10.4.1 Were the students motivated? .................................................................. 328
    10.4.2 Were the students honest? ........................................................................ 328
    10.4.3 Can we extrapolate the results from written assessments to medical student practice and to future clinical practice? .................. 329
  10.5 Implications and future directions ................................................................... 331
    10.5.1 Healthcare professional clinical decision-making practice and education ................................................................................................. 331
    10.5.2 Further research ....................................................................................... 333
  10.6 Closing ............................................................................................................. 334

Abbreviations ........................................................................................................... 335

Associated publications and presentations .............................................................. 337

Chapter 2 .................................................................................................................. 338
  Peer-reviewed publications ................................................................................... 338
  Conference proceedings ....................................................................................... 338
List of Tables

Table 2.1: Performed ability and self-efficacy ......................................................... 34
Table 3.1: Weighting allocation using a standardised total of ten ......................... 75
Table 3.2: Weighting allocation using respondent-derived weightings ................. 75
Table 3.3: Linear balanced scoring system ............................................................. 82
Table 3.4: Logarithmic scoring system .................................................................. 83
Table 3.5: Alternative reporting of data ................................................................. 120
Table 3.6: Self-monitoring for both efficiency and safety ........................................ 123
Table 3.7: Self-monitoring for efficiency but less self-monitoring for safety ....... 124
Table 3.8: Self-monitoring for safety but less for self-monitoring efficiency ....... 124
Table 6.1: Development of papers ............................................................... 166
Table 6.2: Confidence descriptors ....................................................................... 167
Table 6.3: Confidence and safety scoring system .................................................. 170
Table 6.4: Total number (%) of responses by confidence and safeness under confidence and safeness instructions ................................................................. 176
Table 6.5: Internal consistency (Cronbach’s α) for confidence in response, safeness of responses, correctness of response, composite score and number ................................................................................................................. 177
Table 6.6: Number and percentage of responses by year group and instructions .......................................................................................................................... 179
Table 6.7: Number unsafe responses per student .................................................. 181
Table 6.8: Effect of experience on correctness for confidence: % responses correct and % responses incorrect and not unsafe for different levels of confidence and year group ................................................................. 182
Table 6.9: Effect of experience on proportions of all responses that are unsafe for different levels of confidence and year group .......................................................... 183
Table 6.10: Effect of knowledge on % responses correct and % responses incorrect and not unsafe for different levels of confidence and total correct ... 183
Table 6.11: Effect of experience on proportions of incorrect responses that were unsafe for different levels of confidence and year group ...................... 186
Table 6.12: Effect of knowledge on % incorrect responses that were unsafe for different levels of confidence and total correct ................................. 187
Table 6.13: An example of combined confidence (certainty) and safety scoring used to make summative decisions ................................................................. 191
Table 7.1: Certainty descriptors ............................................................. 203
Table 7.2: Responses for students in Score Group 1.............................. 208
Table 7.3: Responses for students in Score Group 2............................ 208
Table 7.4: Responses for students in Score Group 3............................ 209
Table 7.5: Responses for students in Score Group 4............................ 209
Table 7.6: Measures of internal consistency of correct responses and variance in total correct across the 6 tests ................................................. 210
Table 7.7: Odds ratio of a correct response by certainty and score group .... 211
Table 7.8: Odds ratio of an unsafe response among all (correct and incorrect) responses by certainty and score group .................................. 212
Table 7.9: Odds ratio of an unsafe response among incorrect responses by certainty and score group ......................................................... 214
Table 7.10: Odds ratio of a moderately or highly unsafe (clinically significant) response among all (correct and incorrect) responses by certainty and score group ......................................................... 215
Table 7.11: Odds ratio of a moderately or highly unsafe (clinically significant) response among incorrect responses by certainty and score group .. 216
Table 8.1: Certainty descriptors ............................................................. 231
Table 8.2: Total responses for all students ............................................. 238
Table 8.3: Proportion correct and odds ratio of correct for levels of certainty for all ................................................................................................. 239
Table 8.4: Proportion unsafe and odds ratio of any unsafe for levels of certainty for all ................................................................................................. 239
Table 8.5: Proportion and odds by degree of unsafeness among all responses. 240
Table 8.6: Proportion and odds ratio of any unsafe within incorrect for levels of certainty for all ................................................................................................. 240
Table 8.7: Proportion and odds by degree of unsafeness among incorrect responses ................................................................................................. 241
Table 8.8: Proportion and odds ratio correct by gender ............................ 242
Table 8.9: Proportion and odds ratio correct excluding 'don’t know' responses by gender ................................................................................................. 242
Table 8.10: Proportion and odds ratio of certainty of responses by gender .... 242
Table 8.11: Proportion and odds ratio of correct for levels of certainty comparing by gender ................................................................................................. 243
Table 8.12: Proportion and odds ratio for any unsafe response by gender .... 243
Table 8.13: Proportion and odds ratio of unsafe responses by gender..............244
Table 8.14: Proportion and odds ratio of unsafe responses within incorrect responses (excluding correct and ‘don’t know’ responses).................244
Table 8.15: Proportion and odds ratio of any unsafe response within incorrect responses.................................................................244
Table 8.16: Proportion and odds ratio of any unsafe response for levels of certainty comparing by gender...........................................245
Table 8.17: Proportion and odds ratio of any unsafe response within incorrect responses for levels of certainty comparing by gender ..........246
Table 8.18: Correlation for standard assessments results among Test 1 total correct, total unsafe, and proportions of responses for certainty ......248
Table 8.19: Correlation for standard assessment results among Test 1 total correct, total unsafe, and proportions of responses for certainty for Year 2 students.................................................................249
Table 8.20: Correlation for standard assessments among Test 1 total correct, total unsafe, and proportions of responses for Year 3 students ..............250
Table 8.21: Correlation for standard assessments among Test 1 total correct, total unsafe, and proportions of responses for Year 4 students ..............251
Table 8.22: Correlation for standardised assessments among Test 1 total correct, total unsafe, and proportions of responses for Year 5 students......252
Table 8.23: Correlation among Test 1 total correct, total unsafe, and proportions of responses ........................................................................253
Table 8.24: Standard format ........................................................................256
Table 8.25: Free-text box present, with correct response present, option selected ......................................................................................257
Table 8.26: Free-text box present, with correct response present, free-text box selected...........................................................................257
Table 8.27: Free-text box present, with correct response removed, option selected ......................................................................................258
Table 8.28: Free-text box present, with correct response removed, free-text box selected...........................................................................258
Table 8.29: Proportion of cohort response by format/response ......................259
Table 8.30: Proportions of cohort responses with high certainty by format/response outcome.................................................................260
Table 8.31: Comparing responses by standard vs. all non-standard formats......260
Table 8.32: Standard vs. non-standard format, correct options present .......... 261
Table 8.33: Standard vs. non-standard format, correct option removed .......... 262
Table 8.34: Standard vs. non-standard format, option selected .................. 262
Table 8.35: Standard vs. non-standard format, free-text box selected .......... 263
Table 8.36: Non-standard format, correct options present vs. non-standard format, correct options removed .................................................. 263
Table 8.37: Non-standard format, option selected vs. non-standard format, free-text box selected .............................................................. 264
Table 8.38: Test 1 responses of the 19 students who sat both tests .......... 265
Table 8.39: Test 2 responses of the 19 students who sat both tests .......... 265
Table 8.40: Change in responses with time, the difference Test 2 to Test 1 ...... 266
Table 8.41: Proportion of certainty of responses, excluding 'don't know' ........ 266
Table 8.42: Proportion of unsafeness of responses, excluding 'don't know' .... 267
Table 8.43: Proportion of any unsafeness of responses, excluding 'don't know'. 267
Table 8.44: Proportion of any unsafeness among incorrect responses, excluding 'don't know' ................................................................. 267
Table 8.45: Odds ratio correct for levels of certainty .................................. 268
Table 8.46: Odds ratio of any unsafe for levels of certainty for Test 1 vs. Test 2 268
Table 8.47: Odds ratio of any unsafe within incorrect for levels of certainty for Test 1 vs. Test 2 ................................................................. 269
Table 9.1: Certainty descriptors ................................................................. 303
Table 9.2: Internal consistency for responses by year groups for Test 1 and Test 2 ...................................................................................... 307
Table 9.3: Mixed model logistic regression analysis for correct response given factors and interactions between factors ............................................. 308
Table 9.4: Odds Ratios of a correct response relative to below standard students from Year 2 on Test 1 .......................................................... 309
Table 9.5: Proportion of no certainty responses that were correct .................. 310
List of Figures

Figure 2.1: Factors proposed by SCTs as major influence on healthcare professional behaviour .................................................................................................................. 29
Figure 2.2: Context specificity, minimal personal acceptable level and incident occurrence ................................................................................................................................. 38
Figure 2.3: Variable performance over time in different contexts, relative to personal and professionally minimum acceptable levels...................... 39
Figure 2.4: Performance varying with insightfulness .......................................................................................................................................................... 43
Figure 2.5: Relationship of effort and quality of decisions, benefit of self-monitoring ........................................................................................................ 44
Figure 2.6: Representation of self-monitoring for clinical decision-making........ 45
Figure 2.7: Individual’s approach and related ability ‘score’ and error ‘score’.... 48
Figure 3.1: Considering idealness of response and certainty in that response..... 62
Figure 3.2: Certainty by point on triangle........................................................................................................................................................................ 77
Figure 3.3: Logarithmic scoring system .............................................................................. 84
Figure 3.4: Certainty and safe consequences of clinical decisions and judgements ........................................................................................................ 102
Figure 3.5: Classification by correctness and sureness ................................................. 121
Figure 3.6: Interaction of sureness and correctness .................................................... 123
Figure 6.1: Response options .......................................................................................... 171
Figure 8.1: Model to illustrate self-monitoring of item responses in assessment ......................................................................................................................... 288
Chapter 1.  Preface and overview
1.1 Preface

As a mentor helping students to increase their score in assessments, I usually advise students to guess if uncertain. As a clinician supervising doctors at the start of their careers, I advise them that if they are not certain, they should check with me, check with someone else, or look it up. It was this conflict of advice that prompted my interest in this research.

As a clinician and as an organiser of assessments, I know that various responses by students in assessments would have dramatically different impacts if these were made in reality. However, under current testing protocols, it is not possible to know if the student made an incorrect, and potentially unsafe, response as a guess when trying to increase their score, or if this was an answer made with a degree of certainty. That is, we do not know if the student would behave in this fashion and make this choice in an actual clinical setting, or if they would seek assistance. We should know this; it is critical information.

I was already aware of literature regarding safeness of responses when, within a short space of time, I read two publications, one on confidence-based assessment and one on self-monitoring. I drew a link between self-monitoring, a student’s certainty in their assessment responses, and the safety of their responses, and so the basis of this thesis was born.
1.2 Overview

Practicing doctors make diagnostic and therapeutic decisions as part of their daily practice. Doctors require not only sufficient knowledge and skill to make decisions, but also the capacity to know when they have reached the limit of that knowledge and skill, as erroneous decisions can have significant implications for patient welfare. Self-monitoring is this activity of reflection-in-action as part of decision-making. As it is a critical component of practice, medical education needs to assist students to develop self-monitoring.

The next chapter will introduce and review self-monitoring and clinical decision-making. Self-monitoring based in self-efficacy theory relies significantly on an individual’s belief relating to their need for support (or not) in making a decision, based on their perception of their ability to make that decision and the potential consequences of that decision. The importance of knowledge and self-monitoring on effective and safe clinical decision-making is confirmed by an environmental scan of relevant regulatory documentation.

The third chapter briefly explores the purpose of assessment as part of a programme of education. It then explores the assessment of self-efficacy, assessment related to meta-judgements, certainty based assessment and safeness of assessment responses. It concludes with a review of the literature related to assessment of self-monitoring in healthcare, including the openings for this to be investigated further.

The fourth chapter describes the research paradigm to be used through this thesis, and the MBChB course, as the context for medical student education, as delivered during the timeframe of the thesis.

The fifth chapter describes how multiple-choice questions (MCQs) can be developed and evaluated with regard to self-monitoring by including certainty-in and safeness-of item responses.

Therefore this thesis aimed to study decision self-monitoring using certainty-in and safeness-of item responses by investigating the relationships among knowledge, certainty in that knowledge, and the potential consequences of error in decisions.
Chapters six to nine relate to four studies involving multiple-choice tests that included certainty-in and safeness-of responses. In the first study, volunteer students from multiple years in a medical education programme sat a single test having questions in conventional format and in certainty-in and safeness-of response question format. In the second study, all students in a single year sat a single test covering limited content in the certainty-in and safeness-of response format, where the results contributed to progression decisions. In the third study, volunteer students from multiple years sat two tests, with more questions compared to studies one and two, covering broader content and feedback provided in between. In the final study all students sat two progress tests with questions in a certainty-in and safeness-of response format in a single year. These studies evaluated the feasibility of processes for including certainty-in and safeness-of result questions, including potential scoring systems and means of reporting correctness, certainty and safeness of responses. The impact of ability, experience, gender, and associations with other standard assessment formats were variably included. In the third study student perceptions on the approach to, feedback from, and utility of certainty-in and safeness-of response format MCQs was included.

A final chapter includes a discussion of results, linking these to the purpose of the thesis and the underlying framework, and building on the work of others, thereby identifying similarities and differences. The potential value of certainty-in and safeness-of response format is identified, as is some of the work still to be done.
Chapter 2. Introduction to self-monitoring and the concept of self-efficacy
2.1 Chapter outline

Chapter 2 looks at self-monitoring within the various self-reflections and self-assessments emanating from social cognitive theory. It examines self-efficacy theory and self-monitoring as they relate to medical practice, in particular, decision-making and certainty in those decisions.

The first sections, 2.2, 2.3 and 2.3.2, of this chapter introduce self-monitoring, and how an individual’s belief about their capability to undertake actions and their belief about the consequences of those actions are at the core of self-efficacy.

The second sections, 2.3.3 and 2.3.4, apply the concepts of self-efficacy to clinical practice, in that clinicians require not only sufficient knowledge and skills to make decisions, but also the capacity to know when they have reached the limit of their skill and knowledge. This self-monitoring, which is vital to good medical practice, is the activity of reflection-in-action as part of decision-making in daily practice.

The third sections, 2.4 and 2.5, explore the process of self-monitoring diagnostic and therapeutic decision-making, as one of the most important roles of clinicians. Although most of the decision-making by clinicians is successful, a not insignificant number of decisions that individual clinicians make can be incorrect, which can lead to adverse events with significant implications for patient welfare and potential harm. Self-monitoring is one of the strategies used to reduce the frequency and impact of incorrect decisions.

The fourth section, 2.6, is an environmental scan, a purposeful review of several regulatory documents and frameworks for good medical practice. Specifically, this review is used to confirm the place of satisfactory clinical decision-making, the avoidance of harm to patients, and active self-monitoring.
2.2 Self-monitoring

The current model of clinical medical practice, and therefore undergraduate medical education, relies on developing several aspects related to 'self', including self-regulation, self-assessment, and self-monitoring. Accurate self-assessment and self-monitoring are vital to healthcare professionals’ (HCP) self-regulation. Self-monitoring is reflection-in-action as part of daily practice; HCPs need to be aware of when they have reached the limit of their ability or capability. Given the patient and context they are faced with, self-monitoring is manifest when the HCP knows when to “look it up” or “defer to others.” In contrast, self-assessment in this context is a more general judgement of one’s ability. Self-assessment leads clinicians to be aware of the broad content that is required for practice.

Self-monitoring links to the framework of self-efficacy in that it tends to be task-specific, while self-assessment links to the framework of self-concept as it is a judgement of ability that is integrated across dimensions and is independent of context. This distinction between self-monitoring and self-assessment is vital to research, practice, and education. Self-monitoring is the focus of this thesis.
2.3 The concept of self-efficacy

Self-efficacy concerns judgements of how well one can achieve or perform tasks. Self-efficacy contributes to performance by motivating people to succeed, thereby influencing function and behaviour. In this section, social cognitive theories are introduced, and then the implications of self-efficacy theory to healthcare professional behaviour in practice are explored and located within the broader umbrella of social cognitive theories.

2.3.1 Self-efficacy within social cognitive theories

Self-efficacy theory sits within the social cognitive theories (SCTs). SCTs have been used as frameworks to investigate and explain many human behaviours and performances, including those of different healthcare professionals (HCPs) in different contexts. The premise of these SCTs is that people are aware of their intellectual performance, and that awareness influences their decision-making. As shown in Figure 2.1, these theories propose that many factors influence behaviour either directly or indirectly. Specifically, no matter how SCTs are considered, belief about consequences and belief about capability are important influences.

Adapted from, to highlight belief about consequences, and belief about capability as factors.

Figure 2.1: Factors proposed by SCTs as major influence on healthcare professional behaviour
Four human agency concepts that are relevant to SCT informing decisions and behaviour are: intentionality; forethought; self-reactiveness; and self-reflectiveness. They are explained in turn, and related to Figure 1.1, with particular attention being paid to their applicability to HCP.

Intentionality, aligned to intention in Figure 2.1, concerns actions that are planned by an individual. Although behaviours may be intentional, the outcome of a given situation may be completely unintended. As HCP practice is but one component of the complex healthcare system, and system errors do occur, an individual HCP’s practices is not the sole factor responsible for patient outcomes.

Forethought, aligned to the interaction of belief of consequences and capabilities with other factors in Figure 2.1, relates to anticipated events being motivators and regulators of current behaviour. Even with a high opinion of one’s own capabilities, this may not lead to action because of perceived possible undesired consequences.

Self-reactiveness, aligned to the interaction of multiple factors in Figure 2.1, is a multifaceted self-regulation process that links thoughts to actions. It includes the self-monitoring of one’s own behaviour and performance, as well as personal standards and corrective actions such as moral agency. Self-monitoring is reflection-in-action integrated into daily HCP practice, leading an HCP to be aware of when they perceive they have reached the limit of their scope of practice for a given task.

Self-reflectiveness, aligned to the interaction of multiple factors in Figure 2.1, but working backwards from the behaviour, is a retrospective, wide-ranging examination of one’s own performance. For HCPs, this can be considered a self-assessment, a general judgement of one’s performance, and may be used to guide behaviour, performance and future learning at a general level. The distinction between self-monitoring and self-assessment is discussed further (Section 2.4.1).

2.3.2 Self-efficacy theory

Self-efficacy theory concerns task-specific judgements of how well one can achieve or perform tasks. Self-efficacy theory was first developed for analysing behavioural changes in response to psychological treatments. Having been shown to be a useful theory of human behaviour, self-efficacy theory has been
extrapolated widely and accepted as a framework to consider behaviour in various contexts and populations 

Belief about consequences and belief about capabilities are paramount to explaining behaviour. Self-efficacy regulates performance in four major ways: cognition; motivation; affect; and, selection. Cognition relates to one’s judgement of capability and likelihood of success based on one’s knowledge. People motivate themselves and guide their own actions through forethought: they develop beliefs about what they can do and anticipate the likely outcomes of their prospective actions. Affect relates to coping with perceived potential threats and includes strategies such as asking for help, but also includes the potential for doubt to be magnified, causing distress and impaired functioning. With regard to selection, people may avoid activities they believe exceed their capability and coping mechanisms. Even where the activity is required for their role, people may deliberately avoid situations that they feel may exceed their capability or coping 

Self-efficacy for a current task relies on factors including recollection of prior performance and self-efficacy related to that performance. These factors are influenced by four types of experience: enactive attainment; vicarious attainment; verbal persuasion; and physiological state.

Enactive attainment relates to outcomes of personal experience. Prior success, and to a lesser extent, failure, contributes to self-efficacy for a task. People learn from the consequences of their performance. HCPs and HCP students have generally performed well previously and so would be accustomed to having higher levels of achievement.

Vicarious experience is a comparison with others’ experiences. Successful or unsuccessful performance by those perceived to be as competent as oneself, can increase or decrease one’s self-efficacy. HCPs and HCP students would see high achievement amongst their peers and so would be accustomed to high levels of capability. Comparing levels of performance within groups limited to high achievers to elicit differences is difficult, so a perception of common high ability for the group in general may develop, and may be extended to include the self.

Verbal persuasion is a weak contributor to self-efficacy. Perceived support for success from others can reinforce performance. However, the fact that this support
is based on others' observations and interpretations, rather than personal observations and interpretations limits its impact. An appropriate influence on self-efficacy relies on the perception of support, and clear understanding of the task by those providing support. It could be postulated that the learning sequence within many curricula, when applied to healthcare education, introduces learning about, and receiving feedback on, the positive aspects of healthcare (e.g., response to treatments) before the negative aspects are introduced (e.g., side effects from treatments). This inflated view of the likelihood of success can lead to unforeseen consequences (e.g., incorrect prescribing).

Physiological response to stress can reduce perceived capability. Although this is described as being more relevant to physical or practical tasks, clinical environments and clinical learning and associated decision-making can still be considered stressful.

2.3.3 Self-efficacy in healthcare professional practice

HCP practice requires a wide variety of skills as well as the orchestration of these skills to meet the demands of specific situations. Self-efficacy is part of the process of engaging in that orchestration of skills, thus the development of self-efficacy is essential for successful performance in HCP. If an individual does not realise they have reached their limit for a particular situation, the consequences of this could include adverse events and patient harm. Subsequently, if an HCP is unknowingly practicing beyond their limit, their personal development may be misdirected without correction.

HCPs, as they progress with experience and training, differ in their mental representations of clinical problems and decisions. That means that educational programmes need to reflect these differences. As medical training progresses students and trainees engage in activities and decision-making with decreasing levels of supervision. Training progresses to the point where HCPs have appropriate knowledge and certainty such that each has the capability to act independently. For the student or trainee, this is equivalent to, for example, knowing when a supervisor is needed and not needed. This requires personal judgement of required dependence or independence and, with that, self-efficacy.
HCP self-assessment of decision certainty is a component of self-efficacy; and a judgement by the HCP of how certain the decision they are making is the correct one. Clinical decisions are intrinsically associated with an estimate of certainty. Clinicians require not only a considerable amount of knowledge but also an accurate awareness of that knowledge and its gaps. Identifying gaps in ability leads an HCP to take actions to close these gaps, and this happens more when those gaps are identified by the individual rather than through an outside agent.

There are critical times for clinicians when knowledge needs to be applied with a high degree of certainty. Clinicians need to have a strong sense of self-efficacy in order to act appropriately and efficiently in complex or urgent situations that affect the health and well-being of others. Therefore, being able to identify where certainty is lower than it should be is important in any HCP practice and learning. Where the HCP deems that certainty is lower than justified, the HCP may avoid making decisions they actually have the capability to make, demonstrating differential task selection (section 2.3.2). The consequence could be inefficiency or, worse still, detriment.

In the setting of clinicians, self-efficacy can be considered as an individual’s certainty in his/her ability to perform a specific task, including making decisions, in a given domain or context. At some point on the learning trajectory, an individual needs to decide that they are capable of undertaking a clinical task, or when the degree of supervision and/or assistance that is needed can change. When formal certification is required for a specific task (e.g., for advanced cardiac life support), the process of external verification and certification is usually clearly defined to aid the individual’s decisions. Where no formal certification is required, the decision can be left to the individual to make. They usually base their decision on their certainty, background, education and skill and approachability of the supervisor.

It has been shown that personal judgements of ability reflect both a general self-concept and context-specific self-efficacy. These personal judgements of ability can be accurate or inaccurate. One can overestimate or underestimate one’s true ability, and people with low level ability can have a perception of high level ability. Although a potential over-simplification, the intersection between
performance ability and perception of ability can be reduced to a 2x2 table (Table 2.1).

**Table 2.1: Performed ability and self-efficacy**

<table>
<thead>
<tr>
<th>Self-efficacy: perception of ability on task</th>
<th>True performance ability on task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substandard</td>
<td>Appropriately aware of substandard performance</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>Lacking awareness with inefficiency</td>
</tr>
<tr>
<td>Lacking awareness with adverse consequences</td>
<td>Appropriately aware with good function</td>
</tr>
</tbody>
</table>

The ideas in Table 2.1 can be translated into an HCP context. In clinical practice, those misjudging themselves as being of satisfactory ability when not sufficiently able, could make decisions leading to unnecessary treatments and risks of complications. Those with satisfactory ability but without appropriate awareness of their ability will feel the need to seek extra information before making a clinical decision, and this may cause delays in initiating treatment. When awareness is lacking, satisfactory ability with a perception of substandard ability that leads to inefficiency may be more preferable to substandard ability with an overestimation of ability.

### 2.3.4 Self-efficacy and medical education

Given the effect on performance, motivation, and learning, the appropriate development of self-efficacy should be an important part of medical student education and training. It can be expected that individual self-efficacy will evolve with progression though a course of study. Supported by a well-designed course, students need to be provided with opportunities to consider their performance, including their clinical-decision-making ability, to prevent potential for harm to patients.
2.3.5 Summary

In summary, self-efficacy is a task-specific self-belief. Task-specificity is the main feature of self-efficacy that distinguishes it from other motivational constructs and self-belief traits. Therefore, when considering assessment of learning in healthcare context, self-efficacy judgements made by learners need to be made at the task level. Belief about capabilities to undertake actions and belief about consequences of actions are at the core of self-efficacy and social cognitive theories. Self-efficacy requires task-specific cognisance of capability, certainty in that capability, and recognition of the consequences of actions dependent on that capability. Consideration of self-efficacy related to clinical decision-making is vital to good clinical practice, and therefore medical education.
2.4 Self-monitoring for healthcare professions

Self-efficacy is a task specific self-belief based in part on belief on personal capability and consequences of decisions. Although Schön has emphasized the importance of reflection-in-action, specification of achieving active self-observation during everyday clinical tasks is still required.

2.4.1 A research agenda

In 2005, Eva and Regehr asked healthcare professions to reconsider self-monitoring and the associated self-assessment from a conceptual perspective, and develop a prospective research agenda. In discussing the problems highlighted in research, it was postulated that “the roots of the problem in the self-assessment literature involve a failure to effectively conceptualize the nature of self-assessment in daily practice of health care professionals.” Here referring to all the associated self-assessment, but as discussed later self-monitoring and self-assessment are different. This thesis aims to meet some of the need for research related to self-monitoring, as a task-specific judgement, within the context of medical education.

2.4.2 Self-monitoring in clinical practice

Self-monitoring, in a general context, has been described as “self-observation and self-control guided by situational cues to social appropriateness.” Although aligned with this concept, self-monitoring as part of daily healthcare professional practice also relates to clinical appropriateness.

Self-monitoring is essential to good clinical practice. For self-monitoring in clinical settings, in-the-moment cognisance is necessary for the acquisition of data, interpretation, and responses to experiences. In terms of patient safety, self-monitoring is likely to be more important than self-assessment. A clinician’s ability to recognise when a particular situation is beyond the boundaries of their personal competence is likely to be a greater determinant of patient safety than a more generalised awareness of their strengths/weaknesses, and areas for self-improvement. When clinicians make decisions, they need to have an appropriate certainty in being correct. In considering whether the perceived requirements of a task are within, or exceed, perceived readiness to undertake that task, deliberations will include perceived personal capabilities, consequences and
alternative actions. Self-monitoring balances perceived requirements and perceived readiness, and leads to decisions about whether to recruit additional resources. Self-monitoring requires situation-specific cognisance in-the-moment, demonstrating the “interweaving of thinking and doing” required of a reflective practitioner. Paraphrasing Dunning, in order to provide adequate care, a healthcare professional must know when their expertise ends, and the need to call in someone else begins.

For clinicians to adequately self-monitor, they need to be cognisant of some aspect or aspects of their performance in practice, and link this to a standard. In selecting a comparator for one’s own performance, one’s own standards and agreed upon professional standards are relevant. The standard one uses as a comparator could be ‘personal best’, ‘peer performance’, ‘gold standard’, and/or a ‘minimally acceptable’ standard. The standard an HCP uses will depend on the task and the HCP’s personal motivations and goals; their minimally acceptable standard that may be individual or professional. It is imperative for patient safety that any individual minimal acceptable standard be at least the minimal acceptable professional standard.

Figure 2.2 presents one model for demonstrating the journey of improvement in clinical judgement and performance. This model of safe and satisfactory practice, like all models, risks over-simplifying complex situations, but demonstrates the journey for the individual clinician who is continually learning and changing performance. The diagram represents variations in performance over time in practice, with times of improving performance, and times of reducing performance.
In general, unless there is ongoing learning, an HCP’s performance wanes over time; hence the series of downward sloping lines. A learning stimulus can occur, and this may raise self-awareness and help the HCP identify a learning need, which is then filled, and the performance improves: an up-sloping line. If a critical incident occurs and performance unexpectedly falls below the minimum personal acceptable level, this surprise may drive immediate improvement. Alternatively, it is possible that faced with this critical incident the HCP may deny there is a problem, accept the problem but avoid similar situations, or be incapacitated by indecision.

This model assumes that surprise incidents are noticed and acted upon: that when an event occurs or is narrowly averted, it leads to learning and change. The aim is to stay above the minimum acceptable level. These levels will differ between individuals and among tasks. Clinicians may have insightfulness into the problem but fail to undertake corrective learning, or strategically decide it is not needed, as it may be outside their scope of practice and that their minimum acceptable level is lower than that of another HCP with a different scope of practice.

Adapted from “

Figure 2.2: Context specificity, minimal personal acceptable level and incident occurrence

<table>
<thead>
<tr>
<th>Performance</th>
<th>Level of performance for a particular content and context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time in practice</td>
<td>Minimum level for personal acceptability</td>
</tr>
</tbody>
</table>

Surprise event
As illustrated in Figure 2.3, the ability to practice is content- and context-specific, hence performance related to content and context may vary and change at different rates. Practice relative to widely agreed acceptable standards is dependent on context and time. Given content- and context-specificity, an individual clinician will have different levels of ability across content and contexts. In addition, the levels of ability related to these will change, and at different rates and in different directions at any given point in time. There is a minimum personal level for acceptable practice which would be expected to be above a minimum acceptable level within the profession, if not this might warrant scrutiny by a professional body.

Adapted from

*Figure 2.3: Variable performance over time in different contexts, relative to personal and professionally minimum acceptable levels*
2.4.3 Integrating self-monitoring into clinical practice, training and education

Integrating self-monitoring into clinical practice, training, and education requires several qualities on the part of the clinician learner: motivation, attentiveness, and curiosity. A degree of motivation can arise from external sources such as that provided by professional regulations for good practice. These regulations include a requirement for self-monitoring, as will be explained in section 2.6.1.2 through an environmental scan. Internal sources of motivation need to be facilitated. As clinical practice is chaotic and distracting, attention may need to be deliberately directed to developing self-monitoring practices. Finally, self-monitoring requires curiosity. Openly seeking and being comfortable with feedback requires strong self-efficacy, but paradoxically feedback is also most essential at times of insecurity and during states of unknowing.

As an example, experienced surgeons demonstrate self-monitoring by paying more attention and “slowing down” to reduce error during complex intra-operative activity. The interaction between personal action and personal certainty is part of the discussion of safe and efficient clinical practice. Moulton and Epstein described self-monitoring as an important facet of a ‘mindful’ self-regulating surgeon, who is adept at recognising when limits of ability have been reached and there is a need to call for help. However, in focussing on the change of an individual’s routine, this literature does not expand on descriptions of how, when and why individuals call for help as part of mindful self-monitoring.

2.4.4 Learning and developing self-monitoring for clinical practice

There is a paradox within self-efficacy and self-monitoring: one needs to combine faith in oneself with a critical reflectiveness that questions one’s own actions. This delicate balance needs to be developed in students and trainees in environments that are non-threatening, but challenging, and that also emulate clinical practice.

If judgements influencing behaviour reflect both a general self-concept and context specific self-efficacy, how much is due to self-concept, and how much is due to self-efficacy? That is, how much is context-dependent and thus variable? Theoretically, these judgements could be demonstrated as improving (or changing) with learning. Asking students to make a judgement based on self-
concept and self-efficacy, and then giving feedback, may help students to improve such judgements \(^6\). This supports the early introduction of self-monitoring to medical students, accompanied by the provision of sufficient, appropriate feedback. In proposals to develop self-monitoring, learners should not only report their observations, synthesis and decisions related to patients, but also scrutiny of their decisions \(^7,8\). Self-monitoring can be developed from this practice, but this requires the resources to support this learning.

Much of the literature on learning focuses on post hoc reflection-on-action rather than self-monitoring, reflection-in-the-moment \(^8,9\). This post hoc reflection can facilitate self-monitoring \(^10\), however this finding is not universal. Post hoc reflection is usually a general judgement of one’s ability (self-assessment), and does not always imply ability in self-monitoring or vice versa \(^1\). Likewise, encouraging post hoc general self-assessment may not improve self-monitoring \(^13\) or encourage subsequent learning \(^1\).

A failure in self-monitoring may only become apparent when a clinical error or a professional lapse occurs \(^1\). In an effort to reduce errors and/or lapses, an instructional approach to guide learning and future performance that includes assessment focussing on difficult situations in simulations would be useful \(^11,12\). A model of self-monitoring assessment has been described that required relationship-based assessment and a close, trusting and longitudinal relationship between a learner and a teacher \(^\ast\). Implementation of such a model would be resource intensive, both in terms of personnel and time.

As will be expanded upon in Chapter 2, one purpose of assessment is to guide learning \(^\ast\). Such learning could enhance self-monitoring \(^\ast\). Feedback may occur in different forms and formats and at various times, and should relate to both the task and self-monitoring related to that task \(^\ast\). Such feedback will guide ongoing self-monitoring \(^\ast\). The guiding of learning can start before and continue during an assessment, that is, before feedback can occur \(^\ast\). The process of self-monitoring itself may drive learning \(^\ast\).

Epstein \(^\ast\) has developed a general model of self-monitoring behaviour that seems particularly applicable to clinical decision-making. For Epstein, self-monitoring requires certain habits of mind \(^\ast\). The first of these habits is that of novelty: the ability to recognise, consider and describe experiences as being new before
applying judgements, heuristics, or interpretations. Heuristics are personal ‘rules of thumb’ that normally function to aid practice. In a clinical setting, if a situation is not recognised as new and different by the clinician, heuristics might be applied incorrectly, leading to incorrect decisions being made without the usual considerations of self-monitoring.

Conditional referencing is the second habit of mind: the clarification of whether information is primary data, secondary data, or interpretation, and whose interpretation. Recognising differences in data sources and subsequent data interpretation used to inform self-monitoring in clinical practice is an important aspect of conditional thinking. For example, in self-monitoring a decision, a clinician will consider whether they have consulted with a patient personally or whether they are relying on a report from another.

The third habit of mind is considering a situation from multiple perspectives. Considering many possibilities simultaneously without forgetting or excluding any is part of the clinical method of differential diagnoses.

Suspending categorisation and judgement is the fourth habit of mind. Premature categorisation and judgement hinders ability to see a situation as a novel, context-dependent event. Clinicians, and therefore student clinicians and trainees, need not get overly distressed in the face of a new event: they should recognise their personal limits and avoid categorising the event as being equivalent to one experienced before.

Self-questioning is the fifth and final habit of mind. Consciously asking questions of oneself and one’s decisions will help expose incorrect heuristics and biases, and may decrease bias in reasoning.

As will be covered further in section 3.4 for each decision or judgement made, there are associated meta-judgements. Learners (students, trainees) should be able to make accurate meta-judgements in order to prepare for successfully self-regulated practice. The continuum of accuracy for these meta-judgements has been referred to as insight. Accurate self-monitoring makes someone insightful, but insightfulness and performance can vary independently as shown in Figure 2.4.
HCPs are not experts in all aspects of practice, and the process of developing expert thinking takes time and varies across content. In the absence of expert thinking on a topic, self-monitoring one’s decisions is one means of improving the quality of decisions made, as shown in Figure 2.5.
If assessing self-monitoring, this should ideally occur at the same time as assessment of the performance being self-monitored. However, patient and public safety needs to be maintained whilst students, trainees and clinicians are developing their knowledge and skills, including self-monitoring. If authenticity of assessment is reduced, by being less equivalent to actual practice, many of the factors that influence self-monitoring may not be captured. For example, important external factors may be present in actual practice but lost in a simulated environment such as a written test. Therefore, contexts need to be as authentic as possible to actions, while maintaining patient and public safety.

Figure 2.5: Relationship of effort and quality of decisions, benefit of self-monitoring
2.4.5 Self-monitoring in medical degree curricula

Based on the discussion of self-monitoring so far, it might seem clear that self-monitoring should be covered in medical degree curricula. However, it is uncommon to be explicitly included, with only some medical schools now including “self-awareness” as a specified competency. Self-monitoring is a priority for patient safety, which in turn should be a high priority in medical degree curricula.

2.4.6 Summary

Figure 2.6 presents a summary of self-monitoring of clinical decision making. Self-monitoring is the reflection-in-action part of daily practice that leads a clinician to be aware of when they have reached the limit of their capability, knowing when to 'look it up' or 'defer to others', and as such is vital to good clinical practice. Self-monitoring involves judgement about clinical decisions made in the moment; based on one’s belief of capability to make those decisions and the consequence of the decisions. Learning, and development of self-monitoring should be introduced to medical students, undertaken with minimal risk to the students or patients, whilst maintaining clinically authenticity.

Figure 2.6: Representation of self-monitoring for clinical decision-making
2.5 Healthcare professionals’ clinical decision-making

Self-monitoring relates to clinical decisions made as part of daily practice. Making diagnostic and therapeutic decisions related to individual patients is one of the roles, if not the most important role, of healthcare professionals (HCPs) 75.

The process of making diagnostic and therapeutic decisions has been described using many terms and phrases such as clinical reasoning, clinical problem solving, clinical diagnostic reasoning, diagnostic reasoning, diagnostic thinking, diagnostic decision-making, and medical problem solving. The basic process is that the clinician applies their knowledge and experience to information obtained from the consultation with the patient and, via a ‘black box’ of cognitive processing, diagnostic and therapeutic decisions are made. This basic process is well understood, but the ‘black box’ of cognitive processing is a topic of various, and at times competing, theories, models, frameworks and representations 37,75-83.

It is generally accepted that this decision-making process demonstrates case-, content- and context-specificity 37,75,76, whereby an HCP who makes successful decisions for one patient with a set of problems in one context, would not necessarily make similarly successful decisions if these parameters are varied. It is possible that cognitive processing, which is sometimes conscious, sometimes automatic, is different among HCPs for the same task, and among tasks for the same individual 76,77,81,82.

2.5.1 Erroneous decisions by individual clinicians

Although most clinical decisions are successful, or successful enough, erroneous decisions are made 37,76,80. An erroneous decision can occur in diagnosis, treatment, and/or disease prevention 84. There are errors of omission, that is, inadvertently not doing the right thing, and errors of commission, inadvertently doing the wrong thing. This may include subjecting a patient to more diagnostic procedures (acquiring more diagnostic information) than necessary, and/or delaying instigating treatment 85.

An error may or may not cause an adverse event, and not all adverse events are caused by errors 84. Some errors can be serious and even fatal 86. Adverse events are injuries that result from clinical intervention and result in harm. Hospital records analysis estimates that adverse events to patients occur for 10% in-patient
episodes (based on UK, USA and Australia reports ranging from 3.7%-16%). The prevalence for out-patient episodes or those treated in the community by general practitioners is unknown.

Post-mortem studies reveal that 20-40% of post-mortem diagnoses are different to the ante-mortem diagnosis. However, not all of these differences would be of clinical significance. Diagnostic error underlies 10% of adverse events in hospital practice. Diagnostic error is thought to occur at higher rates than medication error and has a worse prognosis, yet is less often reported. It is reported that 74-96% of diagnostic errors include knowledge or cognitive processing factors, these being made by clinicians at all levels and by different means.

Exploration of factors leading to errors has highlight that the individual’s judgements and decisions can be contributory, in addition to system and institutional factors. The lack of knowledge and cognitive processing factors that contribute to the process of making diagnostic and therapeutic decisions errors are potentially numerous.

2.5.2 Individual ability and error

Errors can occur at all levels of HCP ability, and may be due to different factors. Even the most knowledgeable individuals can make errors, including risk-taking and making decisions when not appropriate. An example of inappropriate risk-taking would be incorrectly balancing the benefits and risks of a decision. An example of inappropriate decision-making would be deciding with insufficient information.

Approaches to responding when decisions are required can be different and lead to differing demonstrations of ability, such as correct clinical-decision-making and errors. This is illustrated in Figure 2.7 which shows how individual approaches can be broadly in minimized into four groups depending on ability on tests compared with number of errors of commission. Those with a ‘shot gun’ approach make many decisions, give many responses, and so get a high score on ability, but also make many errors. Those taking a random approach, choose correctly, erroneously, or abstain from making a decision by chance. Those with a constricted approach limit the number of responses they give, but by choosing not to make some decisions, they make fewer errors and get fewer correct choices.
Lastly, those who are thorough and discriminating choose when to respond or not appropriately, and so when they do answer they are significantly more likely to be correct, while not making errors.

Adapted from 85

Figure 2.7: Individual’s approach and related ability ‘score’ and error ‘score’

2.5.3 Institutional, system and individual factors relating to patient safety

Patient safety can be described as avoidance of accidental injury 89, and is affected by institutional, system, and individual factors 87, and so these factors need to be considered in trying to minimize patient risk and maximize patient safety. Many reports and recommendations relating to patient safety take a systems approach 90. This approach is important and useful, but it can be taken at the expense of not considering the impact of individual diagnostic error. Devastating high-profile errors, for example, the removal of an incorrect organ, are rare, whereas diagnostic errors are more common. Missed or delayed diagnosis is an important patient safety problem. Significant diagnostic error rates reported across a variety of clinical contexts generally result in less high profile, but still significant, effects on clinical outcomes.
2.5.4 Individual approaches

Diagnostic error can never be totally eliminated\(^1\), as relevant diagnostic information is often incomplete, inaccurate or unavailable. HCPs should have insight into how personal, institutional and individual factors affect patient safety\(^2\). The use of techniques such as consulting with others and obtaining a second opinion are ways to reduce an individual’s erroneous decisions\(^3\). Acceptance of personal ignorance is preferable to error, so as to avoid unsafe outcomes\(^4\). The combination of misinformation and overconfidence in clinical practice is dangerous\(^5\). Good clinicians can make sound decisions based on incomplete information, but do not guess inappropriately\(^6\).

Medical students often request more information, despite having sufficient information to make a decision\(^7\). Calling for help or seeking assistance for too many decisions, when those decisions are appropriate and within scope of practice, is inefficient. This inefficiency may delay patient management. Some certainty is appropriate because avoiding delay is needed to manage some clinical situations.

Many medical schools now include clinical decision-making as part of their undergraduate curricula\(^8\), but those same medical schools have been slow to include and integrate patient safety in undergraduate curriculum\(^9\).

2.5.5 Summary

An essential role of clinicians as healthcare professionals is making diagnostic and therapeutic decisions. The majority of decisions made by clinicians are successful. However, a not insignificant number of decisions made by individual clinicians at all levels of ability and experience, can be erroneous. These erroneous decisions can lead to adverse events with significant implications for patient safety and potential harm. There are strategies at individual, system and institutional levels to reduce the frequency and impact of erroneous decisions. Education related to these should start with HCP students.
2.6 Validation of self-monitoring as an important part of clinical practice

The attributes that medical students need to demonstrate at graduation are those of a practicing clinician, though the performance level expected in those attributes may be less than that required of expert clinicians. Clinical practice is made up of many components. Self-monitoring is the task-specific and context-specific reflection done in-the-moment as to whether additional assistance and/or resources are required to support decision-making. The tasks considered as part of the self-monitoring process include clinical-decision-making based on knowledge and/or skills. Self-monitoring includes consideration of consequence and need for assistance and, as such, is required for safe practice and avoidance of harm.

To validate the review of the literature to this point, it is important to establish a recognised requirement for self-monitoring of clinical-decision-making as knowing one’s limits, seeking assistance, and safe practice with the avoidance of harm.

2.6.1 Environmental scan

Validating the place of self-monitoring and the associated attributes as requisites for good clinical practice was reviewed as part of a conditional viewing environmental scan. An environmental scan includes looking both at information (viewing) and for information (searching) with the purpose of acquisition and use of information about events, trends, and relationships in the external environment. A conditional scan can be used when most information is gathered from well-established sources, such as formal industry databases and external reports. This method was chosen as there are many documents that describe requirements of clinical practice and clinicians. These requirements specify attributes related to safe clinical practice including clinical-decision-making: which requires knowledge and knowing one’s limits and seeking assistance where present within the varied frameworks and classifications of good clinical practice and medical education.

Formal and well-established documents were chosen to include: regulatory documents both local and international; frameworks for professional activities, which have been widely applied; post-graduate training requirements at
curriculum level and individual outcome level, which have been widely applied; and, lastly, the global minimum requirement of essential education for medical graduates.

2.6.1.1 Extraction from documents as part of environmental scan

The documents were reviewed and the instances were identified related to: clinical decisions based on knowledge and/or skill; seeking assistance at limitations of knowledge and/or skill; and patient safety and harm (Appendix A). These three components were represented within all the documents, apart from there being no specific mention of patient safety, harm or risk within the global minimum requirements.

2.6.1.2 Factors confirmed within documents

Areas that would directly align to self-monitoring, as reflection-in-action as part of daily practice and seeking assistance at the limit of one’s ability, are highlighted here. Good Medical Practice, produced by the Medical Council of New Zealand, includes: “recognise and work within the limits of your competence”; “referring the patient to another practitioner or service when this is in the patient’s best interests”; and “consult and take advice from colleagues when appropriate”. Good Medical Practice, produced by the Australian Medical Council, includes: “recognising and working within the limits of your competence and scope of practice”; “referring a patient to another practitioner when this is in the patient’s best interests”; and “consulting and taking advice from colleagues, when appropriate”. Good Medical Practice, produced by United Kingdom General Medical Council, includes: “you must recognise and work within the limits of your competence”; “refer a patient to another practitioner when this serves the patient’s needs”; and “consult colleagues where appropriate”. The Canadian CANMEDS Physician competency framework includes: “practise medicine within their defined scope of practice and expertise”; “hand over the care of a patient to another health care professional to facilitate continuity of safe patient care”; and “determine when care should be transferred to another physician or health care professional”. The United States Accreditation Council for Graduate Medical Education program requirements include: identify strengths, deficiencies, and limits in one’s knowledge and expertise. The United States
Accreditation Council for Graduate Medical Education outcome project includes: “practice-based learning and improvement identify strengths, deficiencies and limits in one’s knowledge and experience”. Global minimum essential requirements includes: “must demonstrate self-regulation ….. with an awareness of personal limitations including limitations of one’s medical knowledge”.

There are many attributes covered in these documents that do not relate to decisions based on knowledge and/or skill, seeking assistance at limitations and patient safety and harm. However, this environmental scan was conducted to ensure that the attributes given consideration in this thesis were deemed important beyond the research sources used throughout the literature review so far.

2.6.2 Summary

Clinical decision-making based on knowledge and/or skills, self-monitoring and safe practice with avoidance of harm occur as important components across several regulatory documents and frameworks for good clinical practice.
2.7 Summary for Chapter

In this chapter self-monitoring is introduced as a reflection-in-action as part of daily practice with HCPs needing to be aware if they have reached the limit of their ability or capability. The ways in which an individual’s belief about their capabilities to undertake actions, and their belief about the consequences of those actions are at the core of self-efficacy have been demonstrated. In applying this to clinical practice, clinicians require not only sufficient knowledge and skills to make decisions, but also the capacity to know when they have reached the limit of skill and/or knowledge limit. This self-monitoring, vital to good medical practice, is the activity of reflection-in-action as part of decision-making in daily practice. Self-monitoring is one of the core strategies to reduce the frequency and impact of erroneous decisions. In the environmental scan of several regulatory documents and frameworks, self-monitoring is confirmed as a requirement for good medical practice.

Having demonstrated that self-monitoring of clinical decision-making is a core expectation of clinical practice, the practice of self-monitoring needs to be introduced, developed, and assessed within the education of medical students. The next Chapter covers the assessment related to self-monitoring with the context of an HCP training and career and more specifically medical student training.
Chapter 3. Assessment with meta-judgements, item level certainty, item response safeness, and self-monitoring
3.1 Chapter outline

Having demonstrated the importance of self-monitoring being core to good clinical practice, and being based on certainty-in and consequences-of decision-making, Chapter 3 looks at the place for self-monitoring within medical education assessment.

The first section, 3.2, briefly looks to confirm the main purposes of assessment, which include guiding and motivating learning, and making evaluative decisions about the individual learner’s progress.

The next section, 3.3, relates to using self-efficacy and meta-judgements literature to demonstrate how assessment of self-monitoring could be informed by certainty in multiple-choice item responses and safeness of item response. The following sections explore how certainty-in and safeness-of item responses in a written/computer based assessment would contribute to an education and assessment package for self-monitoring. This includes reviewing the importance of authenticity of certainty descriptors and methods of assigning unsafeness classification.

In the last section, 3.4, the literature is explored related to the specific purpose of assessing self-monitoring within the context of medical education.
3.2 Introduction to assessment

The purposes of assessment of healthcare professionals (HCPs) can be considered in several ways. This section includes a brief introduction to two goals of assessment and then a more specific look at the content and format of assessment in relation to self-efficacy and judgements relating to the response that one could make, or has made: both include an estimation of certainty in a response. The literature on assessment of self-efficacy and judgements related to responses will be used to inform a proposed assessment of self-monitoring.

Institutions may use information from assessments to monitor achievement in ways, such as to: rank candidates 106; select candidates 108,109; grade/categorise candidates 106,108; predict future performance 108,107; identify learning needs 63,108,110,111; and, to confirm progress 64. Rather, the two important purposes of assessment relevant to the study reported in this thesis are to inform the decisions on whether performance standards have been met by individual medical students, and to provide information to students to use to guide current and future learning 107.

3.2.1 Purposes of assessment

3.2.1.1 Assessment used by institutions to monitor student performance

Healthcare professional training institutions have a duty detailed in legislation to ensure that those candidates who progress and graduate from a training program are at an appropriate standard 114. Assessment information may be used by the institution to ensure candidates are progressing against criteria for graduation or licensure 63,108,110,111.

3.2.1.2 Assessment used by students to guide current learning and/or motivate learning

‘Assessment drives learning’ is a commonly used phrase, supported by evidence 106,107. Such guidance to learning 67, and/or motivation for learning 63,108,109,117 can come from the information institutions publish in advance of an assessment, such as format, content, scoring and decision making. Although institutions may publish information related to an assessment with a view to guiding and motivating candidate learning, this guidance can be subverted by students’ beliefs about the
importance of the examination or content. This subversion can come from misinterpretation of institutional information due to a variety of factors, such as clarity of language, or from alternative information, such as that from fellow student sources, something referred to as ‘the hidden assessment curriculum’.

'Feedback drives learning', another commonly used phrase, is also supported by evidence. After the assessment, feedback information can be used positively to inform candidates about what they have learnt, to indicate strengths and weaknesses, to guide further learning, and to predict their future performance. Feedback can have no impact, or worse, a negative impact. The nature of impact can be due to the specific information provided, the way it is provided, received or perceived, or a combination of these factors.
3.3 Assessment of self-efficacy

As already discussed, self-efficacy is concerned with judgements of how well one can achieve or perform on a specific task. Historically, self-efficacy measures started as self-judgements of one’s capability to organise and execute courses of action to reach goals, with such judgements expressed as the certainty of performing a task, ranging from 0% to 100%. Subsequently, self-efficacy scales reflected what people believed they could do or what skills they possess for a task, expressed, for example, as success in the defined task. To allow for comparison, the outcomes available should include both the self-efficacy judgement and the corresponding actual capability or performance.

3.3.1 Self-efficacy measurement: comparison of perceived capability and actual performance

The relationship between self-efficacy judgement and performance has been quantified by correlations, congruence, level of attainment, and accuracy. The descriptors used to measure self-efficacy have included numerical scales and a variety of certainty scales and descriptors.

Congruence analysis relates to comparing self-efficacy with true performance. At a simple level, this can be measured by comparing perceptions of dichotomous outcomes of success, or not, with actual success. Matched results would demonstrate congruence, Incongruence is the mismatch between perception and performance. Accuracy, as a comparison of degrees of success between self-perceived performance and actual performance, potentially allows for a more granular analysis than obtained from dichotomous outcomes alone.

Factors that may lead to discrepancies between self-efficacy judgements and performance include: faulty judgement of one’s knowledge; misjudgement of task; unforeseen situation changes; ill-defined measures of self-efficacy; inadequate measures of performance; new experiences changing self-efficacy during course of process; new experiences changing performance during course of process; lack of temporal proximity of the measurement of self-efficacy judgement to the task performance, and a perceived lack of relevance of the judgement or the task; and, predictability of the task based on examples. Self-monitoring judgements should be made when the content is known. As an example, consider the factors involved.
in answering the question “Can you sing the national anthem?”). These could include an accurate estimation of singing ability; an accurate estimation of the required level of singing ability; knowledge of the situation (e.g., whether there is to be a musical accompaniment, whether words will be provided, singing solo or in a crowd, number of people watching); and an appropriate personal measurement and performance measurement for this task (e.g., accuracy of words, tempo, pitch). Even for a simple question, the factors to be considered are many, varied and complex.

In a meta-analysis of self-efficacy and work-based performance, Stajkovic suggested that an inaccurate estimate of task difficulty will lead to an inaccurate self-efficacy judgement. In addition, authenticity of the context of the self-efficacy judgement and performance assessment is important. Self-efficacy is often judged in relationship to real situations, whereas performance may be measured in simulated situations. The simulated situations may not include all relevant environmental elements such as resources (time, staff), interdependence within the organisation, and distractions. Assessment of self-efficacy and performance in authentic work-place based environments is ideal, but can be difficult to achieve. An achievable target might be to incorporate into the medical degree some activities with MCQs that stimulate and develop the concept of self-monitoring. Often, the correlations between self-efficacy judgement and performance is higher in laboratory-based assessment rather than field-based assessment, and probably reflects less impact of uncontrolled variance.

Providing accurate descriptions of tasks, ensuring that means to achieve performance are understood, and ensuring the environment is free of physical and/or psychological stress are important when assessing self-efficacy and performance to reduce factors of irrelevant incongruence, as demonstrated by the national anthem example. Apart from the last, these are important issues to consider in assessment generally, and are discussed later.

3.3.2 Self-efficacy measurement: Generality/Specificity

As already described, self-efficacy judgements are considered to be task-specific rather than generalised. A person can demonstrate self-efficacy for one task, but not for another. Within any domain, self-efficacy can be measured across a narrow range, or more broadly (e.g., item vs. test vs. course). Generally, groups of tasks,
judgements and task-specific judgements of personal ability are different constructs that differ conceptually and psychometrically. Judgements related to more generalised content may not be as useful as judgements related to more specific content, due to weaker predictive value: self-efficacy related to specific tasks is a better predictor of performance than general judgements. Congruence varies between and within activities and individuals, and between levels of attainment in these activities. Turan and colleagues argue that when measuring self-efficacy, judgements should be task-specific. As self-efficacy varies across tasks, it needs to be sampled from each task, similarly to capability. That sampling needs to be sufficient. As the self-efficacy judgement relates to task performance, judgements used to measure self-efficacy need to be adapted to, and relate to, specific tasks such as responses in an assessment. These judgements should reference the performance level that corresponds to the related task.

3.3.3 Self-efficacy measurement: consequences

Not only is it important to include the consequences of decisions, such as the implications of the responses in the assessment, but also the consequences of self-efficacy judgements related to those responses. In reviewing self-efficacy measurement in the workplace, Stajkovic and Luthans highlighted that an assessment will be improved by giving consequence to misjudgements of either self-efficacy or the performance task upon which those judgements were made. Self-efficacy judgements are taken more seriously when the consequences are perceived as significant. If there are no consequences attached to personal performance, then there is less incentive to engage meaningfully in the process. Incorrect self-efficacy judgements will usually have adverse consequences, and this should be reflected in the assessment.

3.3.4 Self-efficacy measurement: assessment providing educational value

According to Bandura, experience improves self-efficacy, learning from experience in a safe environment enables people to learn to calibrate their self-judgement to their performance: calibrating with certainty is the ideal for self-efficacy.

It has been questioned whether self-efficacy is inherent or amenable to educational intervention. If it is amenable to intervention, then feedback from assessment, as
the intervention potentially having the greatest value on education →, should be practicable. Effective feedback, with the opportunity to develop, stimulates self-efficacy which improves subsequent performance →. Feedback should include information to help guide learning and future performance, both regarding the self-efficacy and the task about which self-efficacy judgements are being made.

There is complexity in the interaction among self-efficacy, feedback, and future development. Self-efficacy can influence the response to feedback. Inappropriately low self-efficacy can impair an individual’s ability to improve →, while inappropriately high self-efficacy impairs recognition of need to improve →. If self-belief is inappropriately low or high, then no matter the quality of the feedback, it will not be effective: “I’m so bad, I can’t get better” or “I’m good enough, and don’t need to improve”. Successful functioning is best served by appropriate efficacy →.

The educational environment should challenge learners but not be overly distressing →, so as to allow learners to make mistakes and also ask for help and learn from the experience. Avoidance of potentially stressful activities impedes development of coping skills and results in further lack of competency →.

The safe environment is important for the learners, and given that HCP learning can involve patient care, it is also important that the environment is safe for patients.

3.3.5 Self-efficacy measurement: aggregation

In assessments, information on self-efficacy obtained from certainty-in and safeness-of responses can be aggregated within assessment items, or across items to produce an aggregated result. As Figure 3.1 shows, in considering an item response and certainty. In that response, a single aggregated result aligning to practice may be possible for a decision based on idealness of response (from ideal/correct response to less than ideal/potentially harmful), and certainty.
Figure 3.1: Considering idealness of response and certainty in that response

Aggregating these single outcomes across different assessment items in tests is proposed as being possible in order to generate a composite score. Despite combination scores being described, combining scores from assessments of different domains of HCP practice and self-efficacy is not logical, nor can be justified, given the different domains of capability, certainty and safety. Instead of a single combined score, multiple domain self-efficacy measures could be used to develop a pattern of personal efficacy.
3.4 Assessment of meta-judgements of cognitive capability

There are many terms related to cognition such as knowledge, memory, and comprehension, and for each of these a term with a meta-prefix, such as metacognition is described. Cognition, and the equivalent terms, relates to facts, information, and skills acquired through experience or formal education. Metacognition, and the equivalent terms, relates to a judgement on the first, so that metacognition is a judgement on one’s cognition. Generally, the literature considers cognition and metacognition to be separate, although linked.

Cognition relates the skills to help perform a task, whilst metacognition is the skills used by a person to understand and regulate their own performance. Differentiating those who 'know they know not' from those who 'know not they know not', by assessing metacognition, is critical to HCP making decisions for patients.

There are several things to consider in the assessment of capability and one’s judgements on that demonstration of capability:

- **When** is the judgement made in relation to the demonstration of capability being made?
  - This can be before (predictions) or after (postdictions).
- **How** is the judgement made and how does that relate to how the demonstration of capability is undertaken?
  - This varies, with examples described.
- **What** is the scope of the judgement and demonstration of capability?
  - This is generally related to a specific task or more global overall judgement.

There is extensive literature on assessment and meta-judgements. This section provides examples to illustrate various methods and tools used.

3.4.1 Prediction of global performance made before performance

Pre-test prediction by a learner of the correctness of answers has been used as a measure of metacognitive ability. This global monitoring of potential performance is considered better in higher performers, with higher predicted performance being associated with both higher performance and greater accuracy of predictions. In addition, subjects with judgement of higher potential
performance, gained the most from test feedback, leading to a conclusion that pre-
test predication, as a metacognitive ability, appears to be an important factor in
learning.

Although an association has been found between overall self-predicted test scores
and actual test scores, envisaging the content, difficulty and scoring rubric of a
test is challenging. Candidates making the predictions, and experts developing
and scoring the test, need to have a common understanding of the dimensions
being tested. The more 'global' the attribute, the greater the potential for experts
and candidates to have different understandings. Individual candidates use
personal scales with different personal application of criteria, leading to
differences between performance and judgements. Therefore, a global measure
of performance and candidate estimate of global performance is unlikely to be
useful in the assessment of self-monitoring.

3.4.2 Post-diction of global performance

Good self-assessment on a test is the ability to accurately evaluate the quality of
one's own performance. Post-dictions (judgements after performance) are more
accurate than pre-dictions (judgements before performance). This is not
surprising, as once people have seen and considered the content, they have a more
complete notion of the content their accuracy is being judged against.

If test candidates do not realise they are incorrect, they estimate their score as
being higher than actual, and they can be deemed over-confident.

There are problems with post-diction global 'guess your score' type self-
assessments, both in general, and with regard to clinical practice. Although the
content is known by the candidate making a post-diction estimate, the need for a
global estimate requires the candidate to combine different context-specific tasks
and item responses. As self-efficacy includes a degree of task specificity,
estimates of global performance are unlikely to be useful for assessment of self-
monitoring.

3.4.3 Post-diction of item level performance made after performance: calibration

Calibration analysis is a comparison of judgement on performance (e.g.,
estimate percentage correct) with performance (e.g., actual percentage correct).

One advantage of calibration analysis is that it can also be delivered at an item
level \(^1\), detecting accuracy associated with task-specific ability \(^1\). Estimated percentage correct is commonly used, although an alternative measure is a visual analogue scale \(^2\), a 0-100mm line anchored with 0mm representing 0% accurate, and 100mm as 100% accurate. Therefore, this measure is perceived accuracy.

Investigators have looked for relationships between knowledge and calibration \(^1\). Comparing perceived performance with actual performance allows for estimating over-confidence where the estimated probability correct is greater than the actual probability correct \(^2\). When being tested on tasks on which candidates have no knowledge, with two response options, when proportion correct expected to be near to 50% correct (0.5 probability correct), demonstrate that some people are intrinsically over-confident. For those with some knowledge, training can improve correctness of response, but responses are still over-confident. When candidates were divided into levels of ability by scores, and items by degree of difficulty, the lowest scorers overestimated their scores most on the hardest items \(^2\).

Calibration analyses are useful when the judgement of numerical probability correct can be directly compared to the true outcome of true numerical probability correct. This can allow statistical analysis and interpretation, and can be measured by the correlation between estimated and actual performance \(^1\).

Gamma statistic analysis gauges whether candidates can discriminate between better and worse performance \(^2\). It is an analysis of estimated probability correct for levels of actual correctness, rather than actual correctness for estimated correctness. The important difference between these metrics is covered subsequently in section 3.7.2.2.

### 3.4.4 Miscalibration

Miscalibration occurs when the judgement on performance and actual performance are not aligned. Moore and Healy \(^1\) described three distinct ways in which the research literature has covered this lack of alignment: (a) overestimation of one’s actual performance; (b) overplacement of one’s performance relative to others; and (c) excessive precision in one’s judgement. These can occur independently, for example it is possible to have low score correlation (demonstrating overestimation of one’s actual performance) but have a high
agreement in rank (not demonstrating overplacement of one’s performance relative to others).

Various factors have been postulated as being contributors to miscalibration, including: pressure on students to inflate their marks; belief that the mark awarded by staff is influenced by the self-assessment grade; modesty; socially desirable responses; and inflated self-description. Different factors could lead to miscalibration in opposite directions, with one factor leading to an overestimation in performance and another leading to an under-estimation of performance. Similarly factors could lead to incorrect self-monitoring, as overestimation or under-estimation.

3.4.5 Calibration and education

It has been shown that calibration of judgements on performance and actual performance remain stable over the course of a semester, raising the question of whether such judgements can be taught or learned. Merely prompting students to think about their performance might be too passive a way to alter calibration accuracy, meaning targeted, explicit guidance with practice might be required. It is not just what the learner knows that is important, but also what the learner believes they know. Quoting directly from Van Loon:

To conclude, as Ausubel, Novak, and Hanesian (1978) stated: “The most important factor influencing learning is what the learner already knows. Ascertain this and teach him accordingly” (p. 163). Our findings add to this statement that what the learner believes to know also influences his learning, not only directly, but also indirectly by affecting monitoring and regulation of learning. Therefore, the identification of what children already know, and the detection of inaccuracies in their prior knowledge, might be important to improve their self-regulated learning.

There is a difference between not knowing and ‘knowing’ inaccurately. Over-confidence is higher with respect to commission errors. As knowledge increases, commission errors become apparent, and this is potentially worse than no knowledge. As ability increases, there may be a phase of reduced accuracy of metacognition.
3.4.6 Limitation of some calibration analyses for HCP practice

In clinical practice, the nature of certainty in decisions is different to percentage correct described previously. A clinician will not review a decision on a patient based on whether they are 10, 20,..., 80, 90% correct. For this reason, the model of percentage correct for calibration is not a useful model for self-monitoring of clinical judgements. Problems in the authentic clinical practice are ill-structured, and do not have a single unequivocal solution. Simple problems can have simple right or wrong answers \(^1\), and simple puzzles do not require the consideration of alternative arguments, seeking new evidence, or evaluating the reliability or sources of data. In clinical practice “specific problems are often addressed though consultations with colleagues or with medical literature” \(^2\). The certainty in a clinical decision is not considered as a percentage correct, but the need to seek input from colleagues or literature.

At the extremes of actual performance there are ‘ceiling and floor’ effects \(^3\). If true performance is near the highest or lowest possible, it becomes more likely that an individual will make a judgement that is respectively lower or higher, just by chance.

Metrics used in judgement measurement and performance measurement may demonstrate satisfactory reliability, but when these measurements are combined to produce a measure of metacognition, reliability can decrease \(^4\). Predictions of correctness of responses and performance results produce higher reliability (such as Cronbach’s \(\alpha\)) levels than the actual performance upon which the predictions and certainty judgements are made \(^5\). Predicted scores can be stable despite external factors that could be expected to cause variation, raising the possibility that a global trait such as overarching self-concept might play a significant role in these predictions \(^6\). Some reviewers provide evidence that supports the judgement of performance as a trait, with less task-task variation \(^7\). If such judgements are measured reliably and shown to be stable, this may point to their being a trait \(^8\). These item levels scores may in fact be a reflection of overall perceived ability rather than task-specific ability \(^9\).

Although not specific to calibration analyses, strategies candidates use to answer questions will influence the interpretation of calibration analysis \(^10\). Candidates may not take research assessments seriously, and this will influence results and
therefore the validity of extrapolations drawn from those results \cite{144,145}. Consequence influences motivation; motivation influences performance \cite{146}. The process of calibration with a degree of reward, can lead to candidates attempting to ‘game the system’ \cite{60}.

There are significant concerns over methodological issues in the self-assessment literature. In particular\cite{147}, these include: reliability of a single episode of self-assessment; reliability and validity of ‘gold standard’; a lack of coherence of self-assessment and ‘gold standard’ assessing; use of different scales, problems, and tasks; group level analysis with mean score comparisons; and candidates not taking self-assessment exercises seriously.

In a review of the self-assessment literature \cite{7}, of 5,798 studies only 32 studies were of sufficient academic rigor to include in a review. Problems highlighted included: poor accuracy of self-assessment; unvalidated rating scales; inaccuracy of qualitative measures; a lack of a ‘gold standard’; use of different scales; and student vs. group level analysis. This has led to a call for future research which moves away from quantitative comparison of accuracy, and focuses on distinct aspects of self-concept, self-efficacy, and self-assessments \cite{7}.

3.4.7 **Summary of assessment of self-monitoring as a meta-judgements of cognitive capability**

In summary, extrapolating from self-efficacy and meta-judgement literature, self-monitoring can be aligned to item-level response certainty and safeness, with item certainty descriptors authentic to practice. Scoring methods and score aggregating systems across correctness, certainty and safeness (and incorrectness, uncertainty and unsafeness) need to be developed. Analysis of psychometric qualities, subgroup performance, and individual performance factors, related to decision/judgements and self-monitoring judgements, are required.
### 3.5 Item-level certainty

As noted already, HCP clinical practice, and therefore training, promotes accurate decisions on diagnoses. Alongside this decision-making is appropriate self-monitoring, so that the HCPs appropriately recognise their need for assistance. Certainty based assessment (CBA) would be appropriate to assessment of clinical decision-making, as appropriate certainty is crucial for good clinical practice. It is authentic to assess for certainty when certainty is called for. Although CBA is advocated as being aligned to these purposes and discourages guessing, that claim must include the caveat that CBA is a broad heterogeneous group of assessments with various scales, descriptors, scoring, and additional features. CBA is an examinee judgement method used almost exclusively for multiple-choice question (MCQ) format, and usually refers to examinees expressing some level or degree of belief about the certainty in the correctness of their responses.

This section covers item-level certainty within written/computer assessment rather than exam/topic level certainty or performance assessments (e.g., clinical consultation or procedural skills). It describes how assessments with item-level certainty have developed, and how results for assessments including item-level certainty have been interpreted, applied and received.

#### 3.5.1 Which words?

Many words have been used to refer to ‘certainty’ including confidence, certainty, certitude, certain, sure, surety, sureness, conviction, and assurance. These terms tend to be used interchangeably, and only rarely do authors justify how they might be different. Certain is confidence; certitude is confidence; certitude is certain is sure; sureness is conviction is certitude; confidence is assurance. In this section, unless the authors of cited work specifically use different terms for different meanings, the term certainty will be used to refer to a judgement related to an assessment item response.

#### 3.5.2 Origins of CBA

CBA was first used in the 1930s to improve the reliability of test scores. Two separate authors later referred to 'landmark' articles from the mid 1960s that refreshed interest in CBA. Prior to this, many evaluations of CBA concentrated on reliability as the only outcome and the original aim to reduce
measurement error and improve score reliability continued. Although CBA improves score reliability, other factors, such as testing time and score variance, may alter results, and therefore any improvement in reliability needs to be interpreted with care.

It was postulated that CBA could be used to assess more than just knowledge, and the possibility that CBA might be measuring something new was raised, as a benefit additional to improving score reliability. It was described that CBA scores were better than conventional scores at predicting subsequent CBA and conventional scores. Further analysis revealed that improvements in information provided from test results efficiency could not be explained by improvements in reliability alone, meaning CBA measures a new component as well. Derived scores that reflected a knowledge score and a realism (appropriate certainty in that knowledge) score as independent values were developed. These changes in scoring aimed to reward accurate certainty for levels of knowledge. CBA was not used very widely, being difficult to implement until the development of computer-based responding and scoring; subsequently the potential for its use changed.

3.5.3 Developments of CBA

There are many different applications of CBAs. Given its original development, further work continued to demonstrate that CBA improved statistical reliability and predicted future performance. However, using CBA solely to improve reliability might be at the expense of validation evidence. While improving reliability may be done with the best of intentions, actions taken to do so may reduce the legitimacy of the assessment results. All aspects, psychometric and educational, should be weighed when choosing among scoring methods. How does the scoring method change the inferences that can be drawn from the results? CBA might improve reliability, a function of the increase time required, but at the expense of something else.

Assessment of learning should be directed towards identifying a student’s usable knowledge. Students should benefit from good certainty judgements reflecting the essential skill for efficient study and work, whilst being penalised for misplaced certainty. To this end, CBA has been used to assess knowledge in professional training schools.
CBA may allow for more information about the candidates to be gathered from responses without increasing testing time, such as identifying students giving high certainty incorrect answers, or recognising the continuous variability in degrees of knowledge, which is absent from more conventional dichotomous scoring. A properly functioning CBA should produce responses where the certainty level is dictated by what a candidate knows, and what they believe they know. Candidates may be reluctant to admit incomplete information, however CBA allows them to admit low certainty, ensuring scoring rewards are appropriate.

### 3.5.4 Reported advantages and potential problems of CBA

The reported advantages of CBA format and scoring systems include that it can promote and enhance:

- **Authenticity**, when certainty is important in the work.
- **Deeper understanding and learning**.
- **The discouragement of guessing**.
- **Expression of appropriate certainty**.
- **Feedback for candidates and faculty**.
- **Identification of uncertainties and misconceptions**.
- **Increasing discrimination for the same testing time**.
- **Metacognitive (reflecting on justification) skills**.
- **Reflective learning**.
- **Improved remediation**.
- **Rewards for care and effort in completing responses**.
- **Self-assessment**.
- **Setting tests for candidates with range of abilities**.
- **Student engagement in assessment**.
- **Student experience, with less stress from experience**.
- **Recognition and rewarding of partial knowledge**.

Alongside these advantages, many potential problems have also been reported. These problems with CBA relate to how the responses are scored, and what purpose the results are put to.
There are potential problems with the validity of items. For valuable inferences to be drawn from CBA responses, not only has the certainty judgement to be valid for purpose, but the MCQ items upon which the certainty judgements are based need to be valid for purpose \(^a\). Quality in the development of MCQs is covered in section 5.4.

There are potential problems with the validity of the certainty responses. Concern has been raised that a student who is not certain, as opposed to not appropriately certain, will not do well in such assessments \(^a\). This variability in certainty responding might favour some personality traits \(^a\) and penalise students who are not certain \(^a\). It is postulated that there may be a 'general' certainty factor and an 'item-specific' certainty factor \(^a\). Reduced familiarity in CBA method can influence responses \(^a\), therefore practice CBA questions are recommended \(^a\). As for any assessment, CBA should not be used for high-stakes assessment until students are used to the format \(^a\). It is postulated that the process of CBA itself alters certainty \(^a\). Candidates may give a CBA response trying to strategically increase their score, whilst not answering honestly \(^a\).

There are potential problems with the results. In much of the reported research there is lack of control groups \(^a\). Not all CBA has led to different results compared with conventional scoring, leading to conclusions that CBA adds nothing to conventional scoring \(^a\). Alternatively, the additional resources used by staff and candidates adds little value, and the effect is too small to affect large scale fail-pass decisions \(^a\). Another argument is that certainty is associated with ability \(^a\), so the benefits claimed merely reflect the incorporation of an additional ability score. However, the converse is used to justify the same argument: a lack of agreement between ability and certainty may be due to respondents being good and timid, or poor with no insight \(^a\).

There are potential problems with the generalisability of the research. Most evidence relates to voluntary formative only use \(^a\), with few reporting use for summative decision-making purposes \(^a\). Much of the reported data relates to single cohorts, institutions and/or tests, which may limit generalisability.

Gender difference reporting has been variable. Differences found have been reported as assessment bias, and a lack of difference as no bias; whereas some see these as true results reflecting true differences. Differences due to gender have
been reported as being reduced using CBA scoring compared to conventional scoring \(^1\). However, no gender differences have also been reported \(^2\), and gender difference being present for CBA \(^3\). Lastly, a finding of males being more certain and overconfident on CBA scoring has also been reported \(^4\).

There are potential problems with implementation. Despite the apparent advantages, CBA has not been taken up \(^5\). This may be due to the complexity involved in marking and scoring limiting widespread use \(^6\) prior to computer-based responding and scoring. Increasingly, methods of CBA scoring are complex \(^7\), and even where computer scored, results may be difficult for candidates to comprehend \(^8\). If no additional information is produced compared to number correct scoring, it is not an efficient use of additional resources \(^9\). Novelty can limit implementation \(^10\): this will be true for institutions and/or students, and concerns raised about something new, especially for higher-stakes assessment, will slow implementation.

### 3.5.5 Methods to measure certainty

CBA is based around the assessment of certainty in a response. The question that needs to be asked of candidates is, “How certain are you of your response to this question testing your knowledge?” \(^2\).

There are assumptions made in drawing inferences from, and the requirements underlying, certainty testing \(^2\):

1. The scoring function is appropriate.
2. The consequences of responses are known to candidates. The candidates need to know the choices and pay-offs \(^3\). The pay-off matrix \(^4\) is the interaction of response correctness and certainty.
3. The candidates perform in a rational manner; with a high correlation between knowledge and willingness to express it. Candidates trying to subvert or experiment with the scoring system have been reported \(^5\).
4. The candidates' response strategy is not contaminated by extraneous variables which may bias for or against candidates.

There are different formats of asking certainty, which will be explained:

1. Choose multiple responses, and assign certainty rank items or allocate weighting to reflect certainty.
2. Look at all responses and indicate correct/incorrect (multiple True-False questions) with level of certainty.

3. Select a point within a triangle to reflect certainty in three possible responses.

4. Choose one response and give a level of certainty.

The majority of reported formats relate to the last form, but a brief review of the first three is included, before expanding on the last in detail.

3.5.5.1 Choosing and ranking options to assess certainty

One option is a modification of subset selection, where the candidates select as many responses as they wish to ensure those selected include the correct answer.

Following this they could then rank the selected options according to their degree of certainty, with a relative scoring system. Scoring is relative to the position of the response within the rank. Some options could be ranked equally. There is no penalty for incorrect, but preferentially an incorrect response will push correct response options down the rank. Depending on the scoring formula, once a number of incorrect answers have been scored down the rank, all subsequent correct answers count as zero. This is a complex scoring system given the use of formulas, despite claims of being 'transparent'.

Alternatively, candidates could add a certainty weighting to their selected responses, where they assign more weighting to those items where they have more certainty. Scoring can be by the number correct adjusted for certainty, or a formula score (e.g., correct-incorrect) adjusted for certainty. The weighting allocated by the responder is assumed to be equivalent to their certainty in that response. This does assume that the responders are consistent in their use of weighting. What meaning would this weighting and the inter-candidate and intra-candidate variation have if extrapolated to HCP practice? This would assume that certainty in clinical practice would have some numerical value and in considering options, a numerical ratio could be produced.

For an MCQ with a given number of options, a candidate allocates a number to that response equivalent to weighting of certainty that that answer is correct. This
can be done such that the total weighting is standardized. For example, allocating ten points across items \( \Rightarrow \) (Table 3.1), or allocating proportionally (Table 3.2).

**Table 3.1: Weighting allocation using a standardised total of ten**

<table>
<thead>
<tr>
<th></th>
<th>Allocated certainty that any selected response is correct (total 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td></td>
</tr>
</tbody>
</table>

If C were correct the candidate would score 3/10. The advantage of this is that all question responses are on an equivalent scale. If the candidates were completely certain one response, would they put 10 beside that item and 0 by the others, or if they had no idea, would they put 2 by all five responses.

**Table 3.2: Weighting allocation using respondent-derived weightings**

<table>
<thead>
<tr>
<th></th>
<th>Allocated certainty that any selected response is correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
</tr>
</tbody>
</table>

If B were correct the candidate would score 2/3. The advantage of this is that the candidate develops their own relative weights. The candidate creates their own weighting, and so the scale is personal and potentially not equivalent. If the candidate they were completely certain one response, would they put 1 (or any
number) beside that item and 0 by the others, or if they had no idea, would they put 1 (or at least the same number) by all 5 items.
This method has been developed further, as it can include candidate response subsets that are thought to include the correct answer and/or subsets that exclude the correct answer \(^{152,153,206,208}\). However, this can produce increasingly complex scoring algorithms \(^{10}\) which can be difficult for candidates to understand. If the scoring algorithm (pay-off) is too complex and difficult to understand, this might reduce the validity of the certainty responses.

3.5.5.2 Look at all responses and correct/incorrect (multiple True-False questions) with level of certainty
For each response option, essentially a multiple True-False question, the candidate decides on a scale:

- 5 point scale: statement probably true; statement possibly true; no basis for response; statement is possibly false; statement is probably false \(^{151}\).
- 5 point scale: most likely wrong; probably wrong; maybe right maybe wrong; probably right; most likely right \(^{159}\).
- 5 point scale: true; probably true; insufficient data; probably false; false \(^{185}\).
- 5 point scale: probably true; possibly true; no basis for response; possibly false; probably false \(^{160}\).
- 7 point scale: absolutely sure answer is incorrect; very doubtful answer is correct; somewhat doubtful answer is correct; 50/50 chance answer is correct; somewhat certain answer is correct; very certain answer is correct; absolutely certain answer is correct \(^{185}\).
- 7 point scale words and probabilities: 100 it is certain that; 85-99 it Is almost certain that; 60-84 it is probable that; 40-59 The chances are about even that; 15-39 It is probable that … not….; 1-14 It is almost certain that …not…; 0 It is impossible that …. \(^{151}\).
As an example:

For the following statement, select a response:
Chad is the world’s largest land locked country by surface area.

<table>
<thead>
<tr>
<th>statement is probably true</th>
<th>statement is possibly true</th>
<th>no basis for response</th>
<th>statement is possibly false</th>
<th>statement is probably false</th>
</tr>
</thead>
</table>

As expanded in the section 5.2 relating to general methods, multiple True-False questions are generally not considered the best format of MCQs for clinical decision-making. The scales used can be used in single-best-answer type questions, as reviewed later.

3.5.5.3 Point within a triangle to reflect confidence in three possible responses

Spatial probability measure relates to position within a triangle for a three choice question (Figure 3.2). Each question has a choice of three responses: A, B and C. If the responder is certain of one of these alone is the correct answer, then that response is selected. If they are totally undecided, M is selected. If the responder can exclude C but can’t choose between A and B: with equivalence the responder can mark H; if there is a preference for A, mark G; or a preference for B, mark I.

![Figure 3.2: Certainty by point on triangle](image-url)
There are variations of this graphical response format, but essentially, across all variations, the candidate selects the point that best reflects their certainty across three options. This approach only works for questions with three options.

### 3.5.5.4 Choose one response and give level of certainty

There are many different ways to articulate the certainty meta-judgement, including the use of numbers, words or scales, and with descriptors or not. There is a choice between categorical and numerical scales, subjective probability or a rating scale, and many different certainty scales. An important point is that the candidates need to discriminate between choices, and words and probability have different meanings to different people.

#### 3.5.5.4.1 Probabilistic response

There are considerations, some of which are specific to probabilistic scoring. Students need to be trained to understand relationships between their belief and the probability selected, which must be represented by numerical probabilities. As covered already, there is a significant literature on expressing item level certainty as a perceived probability correct. It is important to differentiate between a probability as the response and a probability as a measure of certainty. Belief can be a probability estimate, and certainty an index of that belief.

Certainty can be 100% in a belief of <100%. For example, my belief on the most prevalent total when two standard dice are rolled a thousand times is seven, and that this will occur in approximately 17% of the rolls, but my certainty in this approaches 100%. The answer to the questions, “What is the total most likely be?” could be seven; “What is the probability that this occur?” could be 15-20%; “What do you estimate the probability that you are correct in this response?” could be 90-95%.

Probabilistic calculations must not be too complex for students and staff to understand. Additionally, there is a need to ensure candidates have time to consider all responses.

As noted previously, probabilistic certainty is not aligned with clinical decision-making. However, some of the assumptions listed for probabilistic certainty in section 3.5.5 are also true for other formats of CBA.
Examples of the different ways that the level of certainty can be applied are now described.

The first example makes use of a Likert scale. Following the selection of the question response, the candidate would select a level of certainty from a Likert scale with descriptors.

There are many different scales in terms of number of responses and descriptors that can be used, including:

- 3 point scale = very confident; fairly confident; not confident.
- 3 point scale = probably wrong; not sure; most likely correct.
- 3 point scale = absolutely confident; not sure; total guess.
- 3 point scale = positive; educated guess; wild guess.
- 3 point scale = absolute confidence; little confidence; random guessing.
- 3 point scale = very sure; fairly sure; not at all sure.
- 3 point scale = accurate; educated guess; random guessing.
- 3 point scale = certain, abstain, uncertain.
- 3 point scale = sure, doubt, guess.
- 3 point scale = certainty, doubt, guess.
- 3 point scale = guess, fairly confident, very confident.
- 3 point scale = confidence high, medium, low.
- 4 point scale = positive; fairly sure; rational guess; no defensible choice.
- 4 point scale = positive; fairly certain; rational guess; no defensible basis for choice.
- 4 point scale = perfectly confident; fairly confident; with little confidence; doubtful.
- 5 point scale = very sure; fairly sure; neutral; unsure; pure guess.
- 5 point scale = extremely unsure; very unsure; somewhat sure; sure; extremely sure.
- 5 point scale = almost a guess; probable guess; neutral; fairly certain; almost certain.
• 5 point scale = very confident; confident; semi-confident; unconfident; not confident at all (guessed).

There are scales with some, but not all, points associated with descriptors:

• 5 point scale: 1 = not sure at all; 3 = somewhat sure; 5 = absolutely sure.
• 8 point scale: 1 = not confident to 8 = highly confident.
• 5 point scale: 1 = random guess to 5 = absolutely sure.
• 5 point scale: 1 = no uncertainty to 5 = a great deal of uncertainty.
• 10 point Likert scale: 10 = very sure, 1 = not sure at all.
• Variable length scales being ranging from 'not sure' to 'sure'.

There are scales that combine words and probabilities:

• 5 point scale: I’m very (90-100%) confident in my answer
  I’m moderately (70-90%) confident...
  I’m somewhat (50-70%) confident...
  I’m slightly (30-50%) confident...
  I’m hardly confident (<30%) confident.
• 9 point scale: 20% just a guess: gave 5 answers and I picked one 1 in 5
  30%, 40% not too sure
  50% could narrow it to a choice of 2;
  60%, 70% one answer seemed better than another, but not
  absolutely certain
  80% very certain, 90%
  100% absolutely certain that I got it right.

There are scales that are based on probability correct:

• Chance of being correct: 50-100%.
• Probability judgement of being correct 80% = correct 8/10.
• Percentage likely correct given four options: 25% (simply guessing) to 100% confident (positive answer is correct).
• Probability correct: 0 – 1.00.
• Odds correct e.g. 2:1 1:1, 1:10, 1:100...
• Dichotomous question probability correct: 0.5 to 1.0.
There are scales based on 'percentage confidence':

- Numerical: 100, 75, 50, 25, 0% confident.
- Confidence: 50-52%, 53-60%, 61-70%, 71-80%, 81-90%, 91-97%, 98-100%.
- “How confident are you that your answer is correct?”: 20% 30% 40% 50%...100%.

There are scales with no descriptors:

- 3 point scale.
- 6 point scale.
- 10 point scale.
- 11 point scale.

There are visual analogue scales (VASs) with no descriptors or anchors:

- VAS with points on a line.

Lastly, there are descriptors of action:

- The closest terminology to self-monitoring was: would be willing to proceed on this basis; need to confirm.

For many of those scales with no descriptors or anchors, the candidates have some guide as to the 'meaning' of the selection through the score weighting. The rationale for not having words is that words mean different things to different people and by awarding marks for each certainty level, so the certainty level is always described in terms of marks awarded. Scoring is discussed later in the chapter.

It has been claimed that probability rather than descriptive certainty has intuitive attractiveness. Probabilities are concise, and express uncertainty more clearly than words. However, this needs an operational definition of probability, which would not be possible for clinical decision-making. A clinician does not consider their clinical decision in terms of the percentage probability it is correct. If using probability, it needs to relate to an event with a confirmable probability. This is not possible for all alternatives in HCP exams.
3.5.6 Methods of scoring based on certainty

The development of CBA has been in part related to the development of scoring systems \(^{171}\), many of which are based on mathematical formulas to reward correctness and accurate certainty. However, a stated aim of CBA is that students should use CBA to calibrate their certainty to their capability, rather than to focus on the marking scheme \(^{166}\), although strategic answering related to marking schemes is likely to confound this intent \(^{211}\).

Postulated principles for scoring of CBA include:

- Scoring should be dependent on accuracy and certainty \(^{189}\).
- Acknowledging uncertainty should score better than guessing \(^{153}\).
- The score should be maximised for an honest degree of certainty \(^{152,159,192}\). Dependent on the scoring system, it may be possible for a candidate with total ignorance to score positively \(^{165}\). However, this assumes validity of composite scoring, which may be an incorrect assumption \(^{127}\).
- The scoring system should reward honest certainty responses rather than the exploitation of the scoring system \(^{165}\).
- Divergent certainty and accuracy is expected to be low \(^{165}\).
- Scores must be appropriately motivating \(^{152,165,166}\).

The scoring schemes are often selected without discussion of the rationale \(^{150}\).

3.5.6.1 Linear balanced scoring

For linear balanced scoring \(^{60}\), the correct and incorrect are balanced, as in the example shown in Table 3.3. For any choice of certainty (called confidence by the author of this Table) the score for correct and incorrect is positive and negative to the same degree.

*Table 3.3: Linear balanced scoring system*

<table>
<thead>
<tr>
<th></th>
<th>Very confident</th>
<th>Somewhat confident</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>+1.0</td>
<td>+0.5</td>
<td>+0.1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-1.0</td>
<td>-0.5</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

From \(^{60}\)
There are many examples of this scoring system as shown in Appendix B.

For this scoring system, if all questions are answered with the same certainty, the final score is the product of the level of certainty score and the difference between the number of correct and incorrect responses. Therefore, assuming the candidate is likely to get the majority of answers correct, the maximum score will be attained from putting the highest level of certainty to all questions 166.

### 3.5.6.2 Logarithmic scoring

A problem with linear balanced scoring is score maximisation with a only the highest level of certainty selected, so other scoring systems have been developed 166. Such scoring functions include equal reward and penalty at lower levels of certainty, but as reward increases, the penalty increases even more 166. This is the basis of logarithmic scoring. In logarithmic scoring of CBA, the differential of the score for correct and incorrect responses is not balanced, nor is it linear 182. An example is shown in Table 3.4 and Figure 3.3.

Table 3.4 shows that the score is maximised by candidates choosing confidence level 3 for questions with a true probability correct (across a number of questions) of 80%, confidence level 2 for questions with a true probability correct 67-80%, and confidence level 1 for questions with a true probability correct of <67%. Therefore, the total score is maximised if a candidate has insight into their true probability correct. The same logarithmic scoring system shown in Figure 3.3 is proposed to be superior to balanced scoring 166, provided it is understood 166. The logarithmic scoring system encourages candidates to answer certainty questions truthfully 166.

**Table 3.4: Logarithmic scoring system**

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score if correct</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Score if incorrect</td>
<td>0</td>
<td>-2</td>
<td>-5</td>
</tr>
<tr>
<td>Probability correct to maximise score</td>
<td>&lt;67%</td>
<td>&gt;67%</td>
<td>&gt;80%</td>
</tr>
<tr>
<td>Odds</td>
<td>&lt;2:1</td>
<td>&gt;2:1</td>
<td>&gt;4:1</td>
</tr>
</tbody>
</table>

from 166,182
Several logarithmic scoring systems are shown in Appendix B.

Different logarithmic scoring function equations will have different pay-offs. These scores that use certainty ratings are still based on mathematical theory. Scoring systems for correct and incorrect results, given the candidate certainty and number of possible options, are potentially complex. The basis is in probabilistic decision theory and logarithmic and quadratic loss calculations. Candidates need to understand the scoring system, and giving them the formula may not be enough. Any system that requires the candidates to understand the marking system more than they need to understand the knowledge being assessed, will undermine the validity of the whole test. If the scoring system is to be used by candidates as intended, then examples of numerical pay-offs have to be understood. Increasingly complex scoring systems have been developed, however simpler schemes are easier to use and also improve reliability, with the simplest scoring functions having the highest reliability. Numerical scores with descriptors of what this means in terms of odds, may enhance numerical analysis and strategic scoring, which might not necessarily be the aim of the system.

Figure 3.3: Logarithmic scoring system
3.5.6.3 Derived scores from probabilistic certainty

Numerical/probabilistic scales and logarithmic scoring allow for the development of sub-scores. The certainty index is calculated by subtracting the actual number correct from the predicted number correct for each level of certainty \(^\ast\). A positive index with higher numbers predicted than actually correct, indicates the candidate is overconfident; a negative index indicates underconfidence. Further to this, more complicated scores for 'knowledge and realism' can be derived \(^{158,165,184,204}\), with an overall composite score made up of knowledge and realism components. The 'knowledge' score is derived from number correct, and the 'realism' score from correct for probability estimate \(^{158}\), with a maximised score where the predicted probability equals the true frequency \(^{178}\). An increasing overall score (combined knowledge and realism scores) may be due to increasing knowledge score and/or increasing realism score \(^{158}\). Realism may remain appropriate, but static, whilst knowledge improves. It is possible to improve only realism \(^{184}\). Reliability analyses reveal that a realism score is reliable and stable irrespective of content \(^{158}\), whereas the knowledge score is less reliable, being more content-specific. Inconsistent results have been described in terms of correlations being present and absent between realism and knowledge scores \(^{158}\). All this implies that these could be significantly different attributes. The addition of a derived predicted correctness for certainty (e.g., realism score, as described above) to a knowledge score to produce an overall score will allow for different interpretations \(^{158}\).

Those candidates who have high knowledge scores and low realism scores have been identified \(^{184}\), and this may be due to lower certainty in correct responses and/or higher certainty in incorrect responses.

Analysis of predicted probability correct for actual probability correct can be developed, and from this, further metrics such as regression equations of predicted for actual probability can be developed \(^{134,158,165,204}\). If the slope of this regression equation is greater than one, the predicted probability correct is greater than the actual probability correct: the person overestimates their ability. Scoring systems related to the regression slope have been developed \(^{134,158,165,178,184,204}\), but the regression must be statistically significant and positive to be meaningful \(^{184}\).
3.5.6.4 Problems with probabilistic certainty scoring

As noted, there are problems with linear balanced scoring systems in that they do not encourage candidates to vary their certainty. Equally, logarithmic scoring has the problem that candidates might not understand the scoring system, but even if they do, probabilistic certainty in responses is not authentic to clinical practice.

The perceived consequences and scoring rules will affect responses. Not all candidates will answer honestly and with a view to maximising their score, whereas others will risk everything to get a high score rather than answering honestly. A tendency to risk-taking will affect certainty levels irrespective of knowledge level. Risk-taking, or the opposite conservative personality, could confound interpretation of knowledge and certainty derived scores. Risk-taking behaviour would be of concern in clinical practice.

If a knowledge score and a realism score can change independently, how can they be combined to produce a meaningful composite total score? The total score could go up or down dependent on the knowledge score going up or down and/or the realism score going up or down, rendering it meaningless.

3.5.6.5 Gamma statistic: certain when correct

The gamma statistic, which is an evaluation of resolution, is calculated as the number of correct decisions made as a proportion of the total. Good resolution is demonstrated by assigning higher certainty when correct. It varies between +1 and -1, with 0 equal to no relationship, with a mean gamma scores of >0 an indication of a degree of discrimination. However, there is an important difference in being certain when correct (gamma statistic) and the other derived metrics based on scores derived from being correct when certain. This aspect of being correct when certain versus being certain when correct is covered in section 3.7.2.2.

3.5.7 Applications of CBA scores

CBA scoring has been used in assessment with various purposes. As alluded to at the beginning of this chapter, the important purposes of assessment include guiding learning and informing progression decisions.
3.5.7.1 CBA: Correct when certain

Generally, there is an increasing likelihood of being correct with increasing certainty. There is considerable variability in the derived calculations from correctness and certainty in terms of how such scores might represent over/under-confidence and calibration. From these there are varied results with underconfidence has been reported as more prevalent than overconfidence, but also that candidates demonstrate being appropriately confident. Overconfidence has been reported especially in responses on difficult tasks, with item difficulty predicting accuracy. Being overconfident in one area is associated with overconfidence in another, irrespective of knowledge, suggesting an underlying confidence trait.

Higher ability students are used to being correct, so tend to have higher certainty. In reports of some research into this, certainty increases more than performance, so overconfidence becomes apparently greater with knowledge increase. However, other studies have found that increased knowledge improves calibration. Further inconsistency in the research includes the finding that question content and candidate ability made no difference to score improvement with CBA derived scores.

3.5.7.2 CBA: score reliability

The inclusion of a CBA score, no matter how derived, aggregated to a knowledge score, can increase the range of total scores, which can increase score reliability. This increase in the range of scores has been interpreted as rewarding high ability students and reducing score improvement due to guessing from lower ability students. However, there are different response patterns by ability. Not all CBA scoring will increase reliability to the same extent. The greatest benefit is for a test with low reliability, as CBA is adding a new variable to the total score, increasing the range of scores, and therefore reliability. If analysis is limited to high certainty question responses, with reduced range of scores, reliability is reduced.
3.5.7.3 CBA: candidate factors

The effect of candidate gender on CBA scoring is variable. CBA scoring has been reported as leading to a reduction in the difference between scores by male and female candidates, compared with conventional scoring. Also reported is that males have higher certainty, but percentage correctness by certainty levels are not different. Others have reported that males exhibit more overconfidence. Such differences can contaminate CBA scoring, but if factors are true to authentic practice, does this invalidate it? Many scales use three or five levels with descriptors for certainty, but most candidates have a limited response range of three or fewer levels.

‘Test anxiety’ is a factor that has received a lot of attention for assessments, written assessments, and MCQs. Students with higher test anxiety are also less likely to have high certainty. Although test anxiety can affect responses, no difference in anxiety has been reported comparing assessments with CBA scoring with conventional number correct scoring.

The origin of the knowledge that is being examined alters certainty in that knowledge, for example, higher levels of certainty have been reported for knowledge acquired via lectures compared with that acquired from textbooks.

3.5.7.4 CBA: predicting future performance

One study showed that there is a low level correlation between CBA scores and subsequent regulatory board exams. However, CBA scores can be useful in predicting subsequent performance when used in conjunction with adaptive testing. CBA scores were better predictors of subsequent final scores in these tests than conventional scores. Using CBA formatively prior to an exam did not improve students’ ability to predict final exam results. Item level certainty may not scale up to the complexity of exam- or topic-level certainty. It has been shown that people are not good at judging their overall performance on a task.

These finding might suggest significant limitations of CBA. However, the aim of this thesis is to use CBA aligned to self-monitoring of clinical decision-making, as opposed to an assessment method with intrinsic purpose.
3.5.7.5 CBA: candidate preference

It has been demonstrated that students indicate a preference for CBA scored exams in evaluation surveys\(^8\). Given a choice in formative assessment, more students choose CBA scoring over normal scoring\(^9\). This preference was most marked for those very close to the pass point, with lower and higher achievers having lower preference.

Students and staff both express concern that CBA requires more time\(^9\), although results have been varied on that count. Completion of CBA did take longer than conventional scoring for some\(^46,104,177,203\), but not all\(^18\).

3.5.7.6 CBA: the effect of CBA on subsequent learning and performance

The use of CBA scoring can lead to improved correct and certainty score in subsequent assessments, a result claimed to reflect improved ability\(^40,104\). Further, CBA scoring itself can lead to a change in behaviour to certainty enough to make independent decisions, demonstrating assessment as learning\(^36,234\). It has been claimed that CBA scoring, as certainty in response, speeds learning, as measured by the time to complete a set task\(^217\).

As noted in section 3.2.1.2, it has been argued that feedback is the most useful learning tool\(^120\) and MCQ feedback is no exception\(^206\). The most useful feedback to students is the kind that guides them to review the content of questions they answered incorrect with the certainty in being correct\(^181\). This directs students’ attention to misunderstood topics.

Some CBA formula scores have been developed such that comparison of the conventional number correct score and the certainty formula score will give a candidate feedback on relative overconfidence or underconfidence as well as feedback on knowledge deficits\(^181\).

3.5.8 Certainty, gambling, and risk-taking

The propensity to take risks varies between individuals\(^211\). For an individual to have an increase in degrees of certainty of being correct, a threshold needs to be reached, called ‘threshold of assurance’\(^152\). A lower threshold would be considered to indicate a greater likelihood to take risks. Response patterns, which may be
linked to personality, lead to consistent or maximal certainty thresholds \(^{25}\). This may be related to candidates being strategic, or less engaged, in the assessment. People are more willing to take risks as certainty increases \(^{28}\). Overconfidence may lead to risk-taking behaviour \(^{29}\), and accepting odds that a well-balanced person would exclude.

Risk acceptability is based on certainty: higher risks are accepted with higher certainty. Risk acceptability in clinical practice will relate to the risk/benefit ratio of therapeutic decisions \(^{25,28}\). Certainty weighting is influenced by risk-taking \(^{164}\). Assessment candidates may use increasing certainty just to increase their score, if higher certainty responses score higher, rather than because this is their true level of certainty \(^{162}\). Analysis of CBA responses has been claimed to provide information on a tendency to gamble when responding \(^{162}\). If assessing this gambling trait is not a purpose of the assessment, this needs to be considered as a potential impediment to validity. The weighted scoring system will have an overall score that will be mostly affected by highest certainty responses \(^{160,162}\). Although a tendency to gamble and/or take risks would be important to recognise in HCP assessment, a aggregated scoring system could confuse interpretation, again highlighting the problems of combining multiple derived scores \(^{173}\). A measure of the risk-taking derived from CBA has been described equating to the number of high certainty incorrect responses divided by the sum of the total incorrect and half the number of 'don’t know' responses \(^{173}\). Various other methods have been used to try to evaluate risk-taking in assessment \(^{173}\), including scoring methods such as: a gambling tendency; correction for guessing; and others with increasing complexity. These methods become difficult to interpret as candidate responses will depend on instructions, motivations and consequences. Risk-taking varies with task, subjective/objective probability, stakes and reward \(^{173}\). In HCP practice, one should not behave as if there is profit to be gained by guessing \(^{152}\). Guessing with no penalty has reward, but no disincentive \(^{173}\). Lastly, in one study individual differences in the tendency to gamble were reported as being associated with a personality factor that affects certainty estimates \(^{174}\), whereas another study found no difference for personality factors \(^{177}\).
3.5.9 Item level certainty and HCP clinical decision-making

Can a learning and diagnostic tool be developed to identify misconceptions and deficits in HCPs’ knowledge? The goal of HCP education is to promote accurate clinical diagnosis and self-monitoring in response to acknowledged deficiencies. It is an aspiration that accurate self-monitoring of knowledge will minimise treatment mistakes, and the pursuit of appropriate professional development to improve clinical competence. Certainty in decisions may be equivalent whether answering a test question or making treatment decisions. Over-confidence and underconfidence have clinical decision-making consequences.

For medical students, underconfidence is reported as being rare, however this may be due to the cohort of candidates, the examination format, or the CBA format. An underconfidence index derived from the CBA responses has been reported to correlate with identification of under-confidence in subsequent intern practice, with no correlation reported for over-confidence. Inappropriate certainty, described as both under-confidence and over-confidence, may be predictors for subsequent problems. Underconfidence may be more noticeable as those who are underconfident may regularly request reassurance, or order too many investigations, whereas overconfidence is only noticed when problems occur.

If CBA responding in education and training does reflect certainty in clinical decision making in practice, then the opportunity for a HCP to identify their own appropriate certainty and inappropriate certainty (be it too much or not enough) in a space without direct impact on patient care would be beneficial. This may allow for appropriate certainty to be re-enforced and inappropriate certainty to be guided.

3.5.10 Evaluation of CBA by students

Students’ opinion of CBA has included that:

- It improves their perceptions of learning.
- Asking confidence before options are answered removes possible cueing; removes students’ perceptions of lucky guessing; reduces guessing; differentiates guessers from others candidates.
• It increases pressure during the test.
• It makes scoring like a game.
• It makes them think more deeply about the question.
• Confident people will score differently.
• Top students will score better.

Students prefer to use CBA formatively. Once students have tried using CBA, they favour continuing with CBA as an assessment format. This is especially the case for those with more higher certainty responses.

3.5.11 Content specificity, judgements and meta-judgements

Medicine involves practice ‘one patient at a time’. Assessment judgements and clinical practice exhibit content specificity. A person’s true ability will vary with the circumstances of the task. Therefore, a person will have variable ability to respond correctly to items in an assessment task, dependent on the subject matter of the item. However, people who are more confident on one task tend to be more confident across other tasks.

How much is certainty affected by a belief of overall ability or a belief of content-specific ability? Analyses of test responses may reflect not only the general certainty of the person, but the certainty for the question response. If true variance is less for certainty than for performance, this may explain why certainty scores have higher reliability than accuracy scores. Factor analysis has identified a ‘confidence’ factor which correlates with, but is distinct from, a knowledge factor.

3.5.12 Non MCQ tests

A randomised trial of MCQs and short answer questions (SAQs) answered with CBA revealed no difference in correct scores. Certainty in correct SAQs tended to be better than for MCQs. Repeat assessment did not improve performance. Appropriate certainty for SAQs was better than for MCQs for questions relating to detail, that is, more specific, narrow content. Certainty in a theme, that is, more general, broad content, even when misplaced, is held more strongly than for detail, again raising the possibility of a ‘general’ certainty level. More certainty in correct than incorrect answers for both formats was found, with most incorrect
answers being held with certainty \(^\circ\). There are those with delusions of higher ability, with continuing belief about their ability, despite evidence to contrary. CBA has been found to be less useful for essay questions \(^\circ\), as students do not tend to write about uncertain knowledge in an essay.

3.5.13 Summary

Item-level certainty in responses in a written/computer-based assessment would be useful in an education and assessment package for self-monitoring. Despite the extensive research on item-level certainty based on probability and/or formulaic score calculation, these cannot be realistically be applied to clinical practice. Descriptors of certainty should be clinically authentic and include action leading to consequence. Analysis of CBA responses based on candidate performance and gender have produced varied results. Analysis of CBA responses and sub-scores has been postulated to indicate a tendency to gamble or take risks, and if such metrics were to be used they should be linked to clinical practice. Lastly analysis of responses suggests item certainty may be based on an overall certainty and content-specific certainty.
3.6 Safeness of responses

When self-monitoring, the consequences of decisions should be considered, and in the case of HCPs, patient safety is a vital consequence. Clinical practice, and therefore training, needs to promote practice centred on safe care for patients. This section covers item-level safeness of responses within written/computer assessments rather than exam/topic-level safeness or performance assessments such as clinical consultation or procedural skills. It describes how safeness of responses has been categorised, and the interpreted and applied within assessments.

3.6.1 Unsafe answers in written assessments

In written assessments, students can make responses that experts believe to be harmful, and from these assessments, a score that is based on analysis of harmful error can increase the validity of the test. A concern common to postgraduate medical boards has been that examinees may have mastered the knowledge required to pass a test without having the judgement to practise safely.

A retrospective analysis of an exam from the US Board of Nephrology identified candidates who achieved the standard to be deemed certified, but gave some responses that would be considered dangerous, so further analysis was undertaken.

Two types of unsafe (called dangerous in this article) answers were defined for analysis:

- Direct, where an incorrect response, if undertaken as an action or lack of action, carries a high probability of serious complication, the risk of death, or undue prolongation of suffering.
- Indirect, where an incorrect response, if undertaken as an action or lack of action which can only be implied, carries a high probability of serious complication, risk death, undue prolongation of suffering.

This United States Board of Nephrology study retrospectively took questions from existing tests and rescored a dangerous answer subtest. Seventy-four of 259 questions had the potential for a candidate to provide a dangerous answer. Within this subset, candidates chose a mean of 7.4 (range 0-19) dangerous answers. Score reliability for the dangerous answer subtest was low (α=0.47), probably due to
there being a lower number of items, and dangerous answers were easy to avoid. Better candidates, as determined by their score correct and prior supervisor reports, chose fewer dangerous answers. Analysis of retrospectively-created sub-exams that were made up of either all questions, questions with dangerous answers or questions without dangerous answers allowed for further analysis. This analysis showed that 6.6% of candidates passed the ‘all questions’ exam but failed the ‘dangerous answer’ exam, while 9.3% passed the ‘questions without dangerous answers’ exam but failed the ‘dangerous answers’ exam.

The discussion of these results raised some issues, including: whether it is unrealistic to expect all candidates to make no errors; the need for a measure of how much institutions should agree on what constitutes a dangerous answer; and how many errors/dangerous answers constituted a fail. However, the instructions for the original exams gave no instructions about avoiding dangerous answers, and so candidate responses may not have been chosen with conviction, but in an attempt to increase their score by making a guess.

Further research developed a definition of a dangerous answer as “an incorrect response that, if put into practice by the candidate, would by omission or commission cause a patient harm”. In this research, a panel of five to six experts reviewed questions, and responses were categorised as: correct; not optimal, not dangerous; and dangerous. Of 903 MCQs, 491 had one or more panelist categorise a possible response as dangerous. The majority of the panel categorised 97 (17%) of the options as dangerous. In all, 8% of questions included a dangerous response, the chance of selecting a dangerous answer was less than random. Of those who passed (number correct above pass mark) the exam overall, 6% gave a dangerous response, compared with those who failed, where 11% gave a dangerous response. Of the candidates who failed the exam made up of all questions, 3.2% would have passed the subset of questions with dangerous responses; of those who passed, 10.4% would have failed the subset of questions that included dangerous questions. Again, the candidates were not informed prospectively that this regrading would be undertaken. The conclusion was that a dangerous answering scoring technique may represent an alternative scoring method, and provide additional information. However, it was acknowledged that some methodological improvements would be required, such as reaching higher
levels of agreement among experts about dangerous options by increasing agreement or increasing the numbers of panel members.

Further concern was that candidates who are 'smart but dangerous' pass exams while having disproportionately high numbers of dangerous answers. In the context of an anaesthetics postgraduate exam, again retrospectively regraded without candidates being informed, 34/175 dangerous answers were identified from a panel review. The definition of a dangerous answer was one identified as such by more than four of eight judges. Surprisingly, some judges categorised the 'correct' answer as dangerous. Results revealed that of the 1036 candidates who passed, the mean number of dangerous answers was 1.6 (range 0-7), compared with the mean 3.4 (range 0-10) of dangerous answers of the 1413 candidates who failed. Further to this, 92 candidates who passed selected four or more dangerous answers, and 631 of those who failed selected four or more. The proportion of incorrect answers that were deemed dangerous was 0.42 for the 92 candidates who passed, and was 0.31 for the 631 candidates who failed (p<0.01).

Passing candidates have fewer incorrect answers, but the proportion of incorrect answers that are dangerous increases. This highlights that, depending on the question format, responses that are not correct can be: don’t know; incorrect, but not unsafe; or incorrect and unsafe. There was no difference in dangerous answers between domestic (American) and foreign candidates. In the accompanying oral exam, there was no difference in provision of dangerous answers comparing those who passed with those who failed, dangerous/incorrect, and comparing those between domestic and foreign and using a comment on the candidate response by a single examiner as “…danger…”, which is less reliable. Again, the conclusion was that secondary (dangerous) scoring might prevent candidates from passing who deserve to fail. Is there evidence that candidates are being passed and practice dangerously? There is no difference in practice or performance in other assessments, so this scoring system may not be useful and may be unjust, as it would fail more candidates, possibly unnecessarily.

A further definition of dangerousness was any act of omission, or commission, in treatment which could result in the patient’s death or injury, that is, the tendency to take undue risk in the care of patients. Is dangerous responding equivalent to risk-taking? The question was asked, could a trait of dangerousness be validated and measured reliably? Two small groups, 'smart but dangerous' and
'dumb but safe', were postulated. A panel of five experts reviewed questions to categorise dangerous answers. Basic science questions were avoided, as these were felt to be unlikely to contain distractors representing dangerous actions. An alternative classification may have considered these as being of indirect danger.

The aforementioned review panel considered commission and omission errors. Greater than 60% of panel agreement was used to define dangerous. Surgery questions (assessed by surgeons) contain more dangerous answers than medical questions. Candidates avoid dangerous answers: the number chosen is less than chance, in total and as a proportion of incorrect. Failing candidates chose more incorrect answers, but the ratio of dangerous/incorrect was not different; the failing candidates simply had more incorrect and more dangerous answers. Question difficulty did not affect dangerous responses, with analysis of the proportion of dangerous to incorrect showing no variance with difficulty (defined by proportion of cohort answering correct). Both passing and failing candidates chose dangerous answers, and the absolute number of dangerous answers chosen was small. The reliability for number of dangerous answers across different subtests given was low. This led to a further question, do we need assessment specifically for dangerous decisions? Rather than a stand-alone exam, performance identification of dangerous performances may require extensive longitudinal tracking.

In a further development content experts considered dangerous answers that would not be indicated for patient care. Non-indicated was classified to nine levels, from one, being incorrect but mostly benign, to nine indicating either highly intrusive, or high chance of morbidity or mortality. This is harder to extrapolate to authentic practice and may depend significantly on the scope of practice of the candidates and health-care system. This research did consider whether dangerous selections remain stable across cases or do they demonstrate content specificity? Building on the nine levels of classification, the risk to a patient was defined as: minimal risk (1-3); some risk (4-6); or risky (7-9). There was at least one non-indicated action response on >40% question responses. Of the 76000 responses only 25 were level 9, and 22 were level 8. Although 671 were level 7, 354 were on a single question. This does support content specificity in error decisions, a pattern of errors found by others. Responses with levels 7, 8,
9 were chosen by passing candidates, with the categories of problems including: misdiagnosis; correct diagnosis but over aggressive diagnosis/treatment; misunderstanding; mismanagement (correct diagnosis but incorrect severity); unexplained. There was a negative correlation between essential positive answers and dangerous answers.

3.6.2 Aggregating unsafe answers into assessment decisions

Information related to unsafe answers needs to be aggregated to inform assessment decisions. Compensation is a process of aggregating several pieces of information with the purpose of building a better picture of performance. The main advantage of combining several pieces of assessment information is an associated improvement in reliability of the measurement: this is acceptable if there is a single trait underpinning these pieces of information. There are two aspects of compensation of results to consider with regard to this: firstly, could or should there be compensation across correct and incorrect and unsafe answers; and secondly, how could or should there be aggregation across different degrees of unsafe answers? Aggregating successful clinical decisions and patients treated with less successful and potentially harmful decisions and patient outcomes to produce an average outcome, would not be consistent with good clinical practice. Regarding the second question, given that errors that occur in practice are not equal and may have different consequences ranging from effectively zero through very little, little, to near-fatal and fatal error and that the more significant errors are rare, and reliability of the measurements that are infrequent is difficult, there is a case that aggregation will be difficult.

Could or should information from correct answers and incorrect unsafe answers be aggregated and compensated? Should the assessment be compensatory, where sufficient correct responses mean that incorrect responses are tolerated, or non-compensatory, where each is considered independently? If compensated, it may effectively mean that the 'average' across patients treated correctly and patients treated incorrectly and unsafely, informs a decision from the assessment, with potentially significant consequences. The dimensionality of the assessment should reflect dimensionality of the construct. In this case, averaging across patients treated appropriately and those treated unsafely is not appropriate for clinical practice, so why would it be appropriate in an assessment? A scoring system
could be developed that reflected both correctness and dangerousness of responses to allow for summation to a single score. An example might be +1 correct, 0 don’t know, -1 incorrect, and -2 unsafe. Although this may seem a good solution and easy to implement, its validity would be questionable. As recorded in at least some studies, as the propensity to select an unsafe answer is not associated with the propensity to give a correct answer, these are different attributes and should not be combined. In addition, combining two independent attributes as an aggregated score will reduce the score reliability.

Could or should there be aggregation across different degrees of unsafe answers? What about just the unsafe answers in assessment being compensatory or non-compensatory where each piece is considered independently? One option that has been tried in open response questions is an automatic score of zero if any dangerous response is given. These dangerous responses may penalise inefficient (too many responses, called maximum error) and killer (fatal to patient) responses equally. In this research, ‘killer errors’ occurred on 17 selected questions: including this scoring option in the analysis increases reliability of scores, therefore it was potentially useful additional information. Maximum error (an error of choosing to do too much) and killer error scores were negatively correlated. Those who chose too many options did not choose killer options. Having said that, the following conclusion was made:

"Although the psychometric evidence supports the use of a single compensatory score, we find it troubling that examinees who make several killer and/or max errors can pass the test through strong performance on other items.”

The compensatory score here was compensating both correct and dangerous and different aspects of dangerous (inefficiency and fatality). Key feature questions include considering the efficiency of investigations, recognising what is critical, and avoiding harm. Typically, there are two to three key features for each patient problem. Scoring key features can include consideration of efficiency and potential harm (e.g., “what five investigations....”), adding one of five marks for each of the correct investigations, but nothing if a harmful investigation is chosen, no matter what else is chosen. The positive and negative factors of investigation selection are different, and aggregating a positive score for good practice with a negative score for harmful practice is not necessarily appropriate.
Validity justification requires analyses of domains of practice to confirm the criticality of skill/error, content representation, and authenticity. However, only including options with dangerous errors that have been recalled as having occurred, although authentic, might limit the options included in the response list.

### 3.6.3 High-stakes decisions based on a low number of unsafe responses

The reliability of an assessment score is affected by several factors, but significantly reduces as the number of items reduces. Given the rarity of unsafe responses, a pattern of error would be required for high-stakes decisions. Weighting of error might constitute a validity threat, as the error might not be one of knowledge but mis-typing or guessing, if the exam format encourages this strategy. If unsafe response are to inform high stakes decisions consideration needs to be given to a number of factors, such as categorisation of unsafe by expert opinion, and assurances made that irrelevant issues such as mis-typing on a single question would not lead changing a fail outcome. Weighting of incorrect responses can be calculated 'objectively' for mathematical questions, but this is more difficult for healthcare scenarios.

Domain-critical error can lead to candidates failing a section of an assessment, and thereby potentially failing the overall assessment. This leads to a requirement of multiple high stakes sections required to achieve an overall pass. A scoring system can produce two running scores: score total and unsafe error total. Most errors are minor, and a candidate with many minor errors might still pass, as long as they meet the overall passing criteria. Critical errors are more likely to lead to a fail. Accumulation of errors can be used to make pass/fail decisions in summative assessment. To inform these high stakes decisions categorising of errors needs HCP input.

What about occasional selection of a dangerous answer by mis-typing or by guessing, if the exam format encourages this strategy? One reason to maintain recognition of unsafe answers, even if decisions do not change, is that the purposes of assessment are more than just making progress decisions. Candidates could potentially learn from their unsafe answers.
3.6.4 Tolerance for error as a learning opportunity

One view of the medical profession is that there is no place for mistakes, and perfection is required. However, error will occur in healthcare and errors will affect patient outcomes. No person or system is perfect. Deviations from optimal acceptable practice will occur, but only some will lead to adverse outcomes and patient harm, ranging from no clinically significant consequence to fatality. Systems and staff training are in place to minimise the frequency and consequences of decisions that are deviations from optimal acceptable practice. Zero deviation from optimal acceptable practice, no adverse outcomes, and no patient harm are overly optimistic, but reducing the degree of harm from adverse events is highly desirable, both by individual and systems approaches. Therefore, tolerance of deviations will vary dependant on the context.

Even if no clinically significant adverse event or patient harm occur, any adverse event should not be ignored, but be considered as a learning opportunity for both the practitioner and the system.

The degree of tolerance of deviations from optimal acceptable practice will vary from one setting to another, so rules for tolerance are required. For example, the level of tolerance for incorrect harmful knowledge would be greater in a student completing an assessment for learning, than in a practicing clinician who was certain that no further consideration or external input was considered necessary. Likewise in clinical practice, tolerance may be higher for an episode of deviation associated with minimally suboptimal management in the short term, as the management decisions would be reviewed at the time of a senior review and the error rectified.

3.6.5 Certainty in unsafe responses being correct

Practicing clinicians should not guess if they have insufficient ability. It is preferable for a clinician to accept their ignorance rather than make an error, thereby reducing unsafe consequences of their patient management decisions. Misinformed decisions and overconfidence when combined in clinical practice is potentially dangerous.

Methods to recognise these aspects in scoring patient scenario questions can include: efficiency; proficiency; errors of omission; errors of commission; or a
composite of these. As covered already, various individual component scores are possible, but composite scores aggregating distinct attributes are not always valid if the combination has no valid meaning \( ^{19} \). However, adding certainty in a response to the safeness of the incorrect response could be valid \( ^{25,26} \), as a highly unsafe decision, with significantly unfavourable consequences held with high certainty would be a significant concern. As shown in Figure 3.4, a correct response may be improved by higher certainty (be more efficient) whereas, for an incorrect response, the combination of adverse consequence and certainty can vary from a near neutral adversity held with low certainty through to a high adverse consequence held with high certainty.

![Figure 3.4: Certainty and safe consequences of clinical decisions and judgements](image)

3.6.6 Summary

Item-level safeness of responses in a written/computer based assessment would be useful in an education and assessment package for self-monitoring. A panel of clinician experts could make judgements regarding degrees of unsafeness given knowledge of the case scenarios and scope of practice of the candidates. No person is perfect, so a degree of tolerance of error used to encourage learning would be needed.
3.7 Specific assessment related to self-monitoring

There is some literature on assessment related to self-monitoring and how this may relate to reflection-in-action as part of daily HCP practice. This can be divided into two types: firstly, when evidence for self-monitoring was inferred from some aspect of the standard examination, for example, response time, changing responses; and secondly, when additional information was collected and used to support evidence for self-monitoring.

3.7.1 Evidence for self-monitoring inferred from assessment responses

Evidence for self-monitoring has been inferred from assessments with question-by-question analysis of: response time; deferring responding/flagging questions; response changing; and a rating on a 100-point scale anchored at extremes for certainty. Several of these studies explored factors other than self-monitoring, and are not reviewed here.

In 2007, Eva and Regehr reported results from a study of undergraduate psychology students. These students underwent a voluntary computer-delivered 60-question general trivia quiz representing six domains of knowledge. In round one, after being presented with a question, the students had the option to answer or pass, the time to do this being the decision time. If there was a choice to answer, there was a free-text box and they had ten seconds to write and submit an answer. This was counted as the response time. In round two, the students had to answer all those that they had initially passed on in the first round.

Apart from proportions correct, no other psychometric data on correctness, such as reliability, were presented. The first analysis related to self-monitoring was 'knowing when to defer', revealing that the round one proportion correct was 65.9%, whereas round two was 4.3% (p<0.001). The second analysis related to self-monitoring was 'slowing down at the borders of competence', with round two responses taking longer than round one. As decision time lengthened, students were less likely to be correct. Responses which were correct had a decision time of 7.1 seconds, whereas those incorrect had a decision time of 11.4 seconds (p<0.001). Correct answers also had a shorter mean response time, namely, 4.5 seconds, compared to the incorrect response time of 5.0 seconds (p<0.05).
The discussion confirmed that participants seemed sensitive, in-the-moment, to their probability of being correct, and would defer appropriately. The discussion goes on that, based on this, it would not be possible to differentiate people into 'good self-monitors' and 'poor self-monitors', because of a high degree of content specificity. This may be true for single episodes of assessment, however when a pattern of self-monitoring is demonstrated over several assessments covering sufficiently different content areas, then it may be appropriate to infer good self-monitoring skills.

Following this study, in 2011, Eva and Regehr, reported results from several linked studies. In study one the results were presented of undergraduate psychology students who underwent a voluntary general knowledge trivia quiz of 60 questions representing six domains. The design was as for the previous study. However, the participants were randomly split into two groups, and those in group one had no certainty rating scale, and those in group two had a 100-point rating scale after each question asking how certainly they had just answered the question correctly. Following round two, the participants had access to a web browser with half of each group being specifically instructed to search for answers to questions.

No reliability data were presented. Again, the students (groups one and two combined) knew when to defer, being 72.6% correct for round one and 6.8% correct for round two (p<0.001). Again as per the 2007 study, the longer the decision time, the less likely the participants were to be correct. Further analysis concentrated on the participants in group two who had the item level certainty ratings to complete. Certainty rating correlated with performance across domains (r=0.81), and certainty rating mapped to decision time. There were no data presented about proportions correct for levels of certainty. In terms of the searching, candidates searched more often regarding questions that they had deferred in round one compared with those they had answered (33.1% vs. 7.9%, p<0.001). In summarising the results, again moment-by-moment self-monitoring is demonstrated, and 'looking it up' did occur for those questions deferred. It was reported that the results related to the certainty rating demonstrated that participants were consciously aware of the likelihood of answering a question correctly. Although this may be true, those explicit data were not presented.
Eva and Regehr went on to report the results of a second study. Study two was identical to study one, except that new questions were used, and questions were either of the free text format or best-from-four-options MCQ format. Results from 90 participants were presented. No reliability data were presented.

The participants chose to answer in round one more frequently for MCQ than SAQ (60.4% v 32.6%, p<0.001). Regarding ‘knowing when to defer’, although there was a significantly different percentage correct between round one and round two, the magnitude of the difference was less for MCQ (49.9% vs. 32.3%, p<0.001) than for SAQ (54.4% vs. 11.7%, p<0.001). Regarding ‘slowing down at the borders of competence’, again although the pattern was broadly similar as previously, the differences in response time and proportion correct were not as great for MCQ question format (round one response time: correct 4.3s vs. incorrect 4.8s, p=NS; round two response time: correct 5.0s vs. incorrect 4.9s, p=NS) compared with SAQ (round one response time: correct 6.6s vs. incorrect 9.8s, p<0.001; round two response time: correct 7.2s vs. incorrect 6.3s, p=0.06). The certainty rating again correlated with the performance across domains for SAQ (r=0.47) and MCQ (r=0.29). Again, on the internet search, questions deferred were more often searched than those answered (23.3% SAQ and 14.4% MCQ vs 6.5% SAQ and 5.2% MCQ, p<0.001).

Eva and Regehr discussed the two studies together. Again, the results confirmed that participants seemed sensitive, in-the-moment, to their probability of being correct, and would defer appropriately: given the correlation with certainty rating, these were thought to be conscious decisions. The reasons for the difference between the SAQ and MCQ format were considered. Could it be the difference between self-generating a correct response and recognising a correct response? Could it be the difference between ‘knowing the answer’ as opposed to ‘recognising the answer’? It was postulated that task-relevant information should be incorporated into the instructions. Regarding the internet searching, this was done by those given explicit instructions and focused on questions that were not answered in the first rounds. Optimistically, it could be surmised that participants were interested to learn; less optimistically, that the rates were relatively low. It was concluded that further research into educational effects is required.

In 2012, McConnell and colleagues reported a retrospective analysis of 3507 examinees from a Medical Council of Canada Qualifying Examination (MCCQE)
Part 1. The computer-based best-of-five-options MCQ component of the MCCQE which was used for this analysis, comprised 196 questions (seven sections of 28 questions; 144 counting, 52 pilot). The selection of items within the sections was adaptive within disciplines. The outcome variables explored were total response time, proportion questions flagged for future consideration, and proportion of questions with initial answer changed (analysed correct/incorrect for initial and final answer). Candidates were categorised as high performers (score > passing score +0.5 SD), borderline performers (score within +0.5 of the passing score), or low performers (score < passing score -0.5 SD).

Again, response time was reported, and questions answered correctly were answered more quickly than those answered incorrectly (56.5s vs. 68.3s, p<0.001). There was an interaction among response time, correctness and performance. Candidates were slower when they were incorrect, and this varied by ability. Analysis of flagging revealed that examinees were more likely to flag questions that were answered incorrectly than correctly (10.0% vs. 5.8%, p<0.001), and again this varied with ability, with high performing candidates more likely (12.4%) to flag than borderline candidates (7.3%), who in turn more likely to flag than low performing candidates (4.2%, p<0.001). All candidates were more likely to flag questions that were subsequently incorrectly than correctly responded to, but the difference was greatest for higher performing candidates, p<0.001.

It was concluded that participants were able to recognise instances where the probability of an error was high, and that the tendency to flag such questions appropriately was greatest for the high performers and smallest for the poor performers. Questions that were initially responded to incorrectly were more likely to be changed (24% vs. 8%, p<0.001), and again this varied with ability. Further, it was concluded the finding that the higher performing candidates were more likely than lower performing candidates to change their responses to questions that were initially answered incorrectly. It was also argued that this suggests that higher performers were more aware, in-the-moment, of when they were making an error in their initial response to a particular question.

Additional analysis included a reliability analysis, using the difference between the percentage correct and percentage incorrect, as the 'score', allowing for response time, flagging and response changing, and then analysed for each
content domain and then across the six content domains. Response time reliability index (0.38 for six-item reliability) was relatively poor compared with the other indices for question flagging (0.73), response change (0.79), and test score (0.88).

In discussing their findings, it was first concluded that candidates, when likely to make an erroneous response, were more likely to take additional time, more likely to flag the question, and more likely to alter the initial response. Secondly, these indices are modulated by ability, as defined relative to a passing score. Thirdly, some of the indices of self-monitoring were reliable in this context. Response changing may be an indicator of self-monitoring, with changed responses being more likely to be incorrect originally. However, if the responses are more likely to be incorrect (best-of-five-option MCQs), then it is more likely by chance that for a group of questions answered and changed randomly, more changed responses will be incorrect. For every twenty answered and changed at random, twelve will be changed from incorrect to incorrect, four changed correct to incorrect, four incorrect to correct. The chance of changing responses from correct to incorrect is the same as incorrect to correct, but the chance of change from incorrect to either correct or incorrect is much greater than from correct to incorrect. Personal communication with corresponding author [Assistant Professor Meghan McConnell, Department of Innovation in Medical Education, University of Ottawa Canada, 27 October 2016] explains:

“we went back and forth on how to analyze the response changing data. Ultimately, we decided to go with the accuracy of their first response because, at least theoretically, these would be instances where we would clearly see self-monitoring behavior. So if they initially chose the incorrect responses, were they more likely to change their response.”

Agrawal and colleagues, report a study of 67 final year medical students. The participants undertook a computer based 60 best-of-four-options MCQ (ten from each of six clinical domains) test. The initial test used a correction factor of incorrect answers. On presentation of the stem, the participants had to decide whether to attempt the question or defer (time to attempt defer decision). The decision to answer submission time was taken as the answer time. Following this they responded on a certainty scale (0 = no certainty, 1 = total certainty). Deferred questions were re-presented with no correction factor in place. Participants gave their best answer and confidence.
There was no difference in responses by gender. No psychometric analysis, such as reliability was reported. Despite candidates attempting fifty-five of the sixty questions, the corrected rate for attempted questions was higher than for deferred (71% vs. 40%, p<0.001). Decision times were higher for vignette-based than fact-based items, postulated to be due to item length. Of the questions attempted, those answered correctly were answered faster than those answered incorrectly (10.8s vs. 13.5s, p<0.001). A pattern of reduced likelihood correct with increasing decision time of attempted items was found. Candidates assigned higher mean certainty to questions answered correctly compared with those answered incorrectly (70.1 vs. 46.0, p<0.001).

In discussing their results, it was again noted that participants slowed down when they had doubt in their own ability.

3.7.2 Assessment of individuals for evidence for their self-monitoring

3.7.2.1 Assessment including item-level certainty

Certainty based scoring in medical/HPC student assessments has been investigated on several occasions, but usually within a framework of score analysis, calibration, and/or psychometrics. This section will concentrate on studies with a framework aligning to self-monitoring, or something akin to self-monitoring, in practice.

In 1993, Mann described using item-level certainty, framing it as 'diagnostic metacognition', the practice of seeking additional information when 'underconfidence in a correct diagnosis' but not wasting time, effort and money, whilst a patient's condition could worsen. Mann pointed out that calibration analysis where the 'physician is correct on 70% of diagnoses in which he or she expressed 70% confidence' is of little use. He went on to postulate a better analysis of adjusted resolution, which ranges from zero to one, and is higher when different certainty levels to correct versus incorrect diagnoses. Forty-seven medical students undertook a voluntary test, the content related to cardiac dysrhythmias and ECGs. Students graded certainty on an eleven-point scale (from 0-100%). Although not stated, this probably equates to perceived likelihood correct, as subsequently calibration curves were produced.
No reliability measures of score or certainty level were provided \cite{257}. Although it is stated that candidates were slightly underconfident, the mean accuracy of 85.1%, with a mean certainty of 82.3%, and no statistical analysis of this difference was provided, as adjusted resolution was preferred. The adjusted resolution analysis found that, in general, the level of certainty did not provide information on likelihood of being correct, although this did vary considerably between subjects. The highest levels of resolution were achieved at approximately 85% correctness and certainty.

In discussion, Mann accepted that the content of the test was very narrow \cite{257}. The prevalence of underconfidence was felt to be due to the items being easier than the subjects expected. The highest level of correct and certainty of 85% was potentially being an artefact of the method. For items with true 100% correct, there may be a ceiling effect whereby subjects can only be correct with underestimated certainty. The measure and analysis used in this research is a measure of being correct or not for different levels of certainty. The importance of this, as opposed to being certain when correct, is discussed subsequently.

In 2005, Friedman and colleagues \cite{258}, highlighted the importance of personal knowledge and also the awareness of limits of personal knowledge and use of external knowledge and support. This set the scene for an exploration of item-level certainty (called confidence). For a set of case scenarios, 216 subjects (medical students, residents and internists) were asked to review the cases and develop a differential diagnosis list of up to six possibilities. For each of their options, they responded how likely they would be to seek assistance on a four-point scale from unlikely through to likely. Certainty in being correct was taken to be the reciprocal of seeking assistance. The process was repeated with decision support available. For each subject they calculated the threshold level of certainty to distinguish incorrect from correct. This varied between subjects with maximising correctness varying between those for certainty levels one, one and two, and one, two and three. Using responses with certainty of one to indicate high confidence, and four to indicate low confidence, analysis demonstrated differences by experience.

There are no reliability analyses provided for correctness or certainty, but kappa agreement of correctness with certainty was provided between 0.2-0.4 \cite{258}. There was a significant alignment of certainty with correctness across all subjects,
between correctness and training and certainty and training (p<0.001).
Proportion correct for students was 64.3% for high certainty responses and 19.1% for low certainty responses. Proportion correct for resident doctors was 58.8% for high certainty responses and 35.3% for low certainty responses. Proportion correct for faculty was 63.5% for high certainty responses and 29.4% for low certainty responses. Students (with the lowest levels of training) had the lowest proportion incorrect for low certainty responses. When correctness and certainty did not align, this was more likely to occur in the direction of underconfidence, with the students being more underconfident than the residents.

The findings were discussed as a demonstration of idiosyncratic interpretation of the certainty scale. It was postulated that the greater level of alignment was due to the difficulty of the cases and low proportions correct. It could not be concluded that students were generally more aware, just that they were more aware on these difficult cases. In the discussion it was reinforced that the certainty scale is grounded in clinical behaviour, and that this is strength of the study.

In 2008, Butler and colleagues, investigated the effect of feedback on correct answers and the interaction of certainty in those responses. The framework for this was one of monitoring the accuracy of judgements as a certainty level, and the role this has on efficacy of feedback, especially considering low certainty correct answers. The importance of feedback on high-certainty errors was described, but the effect of improving certainty for low-certainty correct as an area for investigation. In Experiment 1, 30 psychology students sat a 40 question general knowledge test comprising 40 best-of-four-options MCQs presented in random order, with forced response and certainty level on general knowledge question. Feedback on half the questions was provided, and then the subjects sat a second test. The certainty scale was four point (guess, low, medium, and high confidence). Following a short delay, the subjects sat a 60 question MCQ test, without forced response, which included the original 40 question test (for which students had received feedback on 20 of the questions). In Experiment 2, again there was a 40 MCQ general knowledge test with feedback, followed by a delayed test. The differences were that the certainty was on a numerical scale 25-100% (25%=no confidence) for the first test, and 0-100% (0%=guessing) for the follow-up test. The follow-up test was delayed by two days, rather than a matter of minutes.
In both Experiment 1 and Experiment 2, the score on the delay test was increased by the presence of the questions on the earlier tests, and more so by feedback (Experiment 1, proportion correct 24% increased to 87% with feedback; Experiment 2, proportion correct 33% increased to 83% with feedback). Feedback increased the number of incorrect responses that became correct, and also to ensure that correct responses remained correct across both experiments.

In Experiment 1, when the initial response was incorrect, the proportion correct in the final test was significantly improved by feedback (Experiment 1, proportion correct 3% increased to 82%; Experiment 2, 9% increased to 73%, p<0.001), and especially so for high certainty responses (Experiment 1, proportion correct 79% increased to 93%; Experiment 2, 78% increased to 93%, p<0.001). This was linked to a previously described ‘hypercorrection effect’. Feedback also increased final test correct responses. Again there was an interaction with feedback and certainty, with all levels of certainty except high, demonstrating an improvement.

In Experiment 2, the final test was delivered two days later. Here the hypercorrection effect appears to have been replaced by hypocorrection: the benefit of feedback was reduced for high certainty responses. The provision of feedback led to an improvement in gamma correlation (improved by being more confident when correct), with almost perfect correlation of 0.94.

The discussion of the results included that in the absence of feedback, incorrect results are unlikely to be corrected spontaneously; correct responses benefitted from feedback, with a benefit greatest for low-confidence responses; and feedback improved resolution and calibration. The new findings were that low confidence correct responses were subsequently improved on recall, although this may have been related to methodological differences in this and prior studies, specifically the use of forced responding as opposed to free responding. Low-confidence responses may not be included when responding is free, and the greatest benefit of feedback for correct response was seen for low confidence responses.

The second new finding was that feedback enhanced the accuracy of confidence judgements: subjects were more likely to be increasingly confident when correct. Lastly, the hypercorrection effect, causing higher confidence errors to be more likely to be corrected, was only a transient effect, and such correction may not be seen over longer term. Feedback should not just relate to incorrect answers, but to both correct and incorrect, and relative confidence. Formula scoring may impact
on lower confidence responding, and therefore reduce the potential benefit of feedback. The gamma statistic which, as will be discussed subsequently, improved by being more confident when correct.

In 2014, McMahan and colleagues \(^{260}\), framed the purpose of testing as being to estimate what individuals know, but also what they do not know. Admitting what one does not know is an essential element of HCP learning and practice. The dissociation between this and the usual rules of encouraging guessing and associated reward in test-taking included. As stated, “A willingness to openly admit deficiencies is an essential part of …skills required for lifelong learning skills”. A scoring algorithm was created where an incorrect response incurred a negative score, with a positive score (less than that for a correct score) for an unanswered question.

Dental students were informed of the formula scoring, but not the various formulas, and sat an MCQ and SAQ test \(^{260}\). It was demonstrated how varying the incorrect and unanswered score can lead to different expected scores for different levels of ability and guessing, and how these variations can improve the correlation with SAQ scores.

In the discussion of implementation, the impact on students needing to consider the scoring system was raised \(^{260}\), in terms of response time and time pressure within the test. Also, it was pointed out that the students need to understand the probabilistic nature of the scoring system in terms of risk and reward. It was postulated that this might be improved by presenting some example probabilities for students.

Although this study was framed within 'knowing what you know and don’t know', 'not guessing' and 'continued personal development', probabilistic and a variation on formula scoring was used try to achieve this goal \(^{260}\). The decision to guess answers or leave questions unanswered is made on knowledge of scoring probabilities rather than any of the considerations that occur in clinical practice.

In 2016, Celilio-Ferandes and colleagues \(^{261}\), in addition to considering cognitive processes in the application of knowledge, set out to explore awareness of knowledge and metacognitive knowledge. This was framed in the authenticity of acknowledging what an individual does not know given that they will make high-
stakes decisions about patients. This was achieved by analysing results related to the use of “I don’t know”, which is, effectively, “I am not certain enough to answer”.

Data for this study were taken retrospectively from the records of the Dutch progress test, which is a formula-scored progress test \( ^{26} \). A parameter of ‘judgement of knowledge accuracy’ was derived as the number of ‘don’t know’ responses divided by the sum of ‘don’t know’ responses and incorrect responses. Data were available for 347 pre-clinical year and 196 clinical year students.

Questions were categorised by two researchers as being simple (the level being remembered or basic understanding) or vignette (the level being applied, analysed, evaluated, existing knowledge).

The judgement of knowledge accuracy reduced between the first and last years for both groups of students (preclinical, judgement of knowledge accuracy 0.92 reduced to 0.52, \( p<0.017 \); clinical, 0.71 reduced to 0.37, \( p<0.017 \)), and this was true for both question types \( ^{26} \).

These results were discussed, in that although the students got more correct as they progressed, they also got more incorrect \( ^{26} \). This was inconsistent with the finding by others, that those with higher knowledge had higher metacognitive ability. Students in later years were less accurate. The causes of this are postulated as possibly being related to the benefit/risk given the formula scoring.

If the penalty was not high enough, risk-taking behaviours may have increased. Another postulated reason is that as students progress through the course, they are expected to score higher, and may be overconfident due to their increased experience.

The limitations were acknowledged as being research in a single-institution, retrospective analysis, and the potential confounding caused by formula scoring and students’ approach to this \( ^{26} \). In considering implications, it is possible that responding to questions is less effected by judgement of knowledge, especially in later years, perhaps because of adapting answering strategies.

In 2017, Ryan and colleagues \( ^{26} \) investigated the effect of feedback on several metrics, including response certainty, within a framework of self-regulation and self-monitoring. The research question related to how different types of feedback,
including clinical presentations, clinical presentations with high certainty incorrect highlighted, and normative, might affect subsequent test performance.

Students from year two of a medical school sat four progress tests in the course of the year 262. The tests were randomised to different feedback formats. The main results related to the scores on progress tests and summative end-of-year exams, allowing for feedback group and score on Test 1. Item-level certainty consisted of low, moderate, and high certitude.

Students randomly assigned to normative feedback did less well that those randomly assigned to clinical presentation feedback (with or without certainty levels) especially those who had the lowest scores on Test 1 262. Although there was no statistically significant relationship between correctness and certainty within any feedback group, the odds of providing correct responses was highest for responses with high certainty, and increased from Test 1 to Test 4.

As this study investigated a much wider topic (feedback) and the certainty in item responses was a relatively small part, this was not covered in the discussion 263. Either this study was not able to detect any benefit to subsequent performance by including response certainty in feedback, or no such benefit exists.

In 2018, Grazziotin-Soares and colleagues 263 framed their study in identifying students’ misconceptions, incorrect knowledge held as true. Misconceptions are difficult to identify and resistant to change. Measuring confidence and correctness was put forward as a means to identify misconceptions. This is authentic to practice, as calibration between correctness and confidence is an important attribute for clinical practice. A student who is incorrect and not confident is open to learning. In addition, a student who is correct and not confident is guessing, a point that may go unnoticed in traditional assessment. Lastly, a student who is incorrect and confident is misinformed, is less likely to seek additional opinions, and is unaware of their responsibility for adverse outcomes. The setting of this study was third year dental students sitting two assessments, each of 20 best-of-four-options MCQs on endodontics and dental implants comprising ten scenario-based questions and ten factual questions. Two staff independently categorised questions as scenario-based or factual, and agreed by consensus. Response confidence was very sure, sure, unsure, very unsure, dichotomised to
confident or not confident. Students were informed that confidence was important, but would not affect grades.

The results revealed that for scenario-based questions on endodontics, there was a mean of 9.7 misconceptions (incorrect confident) from 105 answers, and 5.4 for factual knowledge questions. Mean confidence (presumably the proportion of responses confident) was 88.5% for scenario-based questions and 94% for factual knowledge questions. For scenario-based questions on dental implants, there was a mean of 22 misconceptions (incorrect confident) from 104 answers, and 23.2 for knowledge questions. Mean confidence was 82.6% for scenario-based questions and 83.4% for factual knowledge questions. Misconceptions were more likely for scenario-based questions related to endodontics, with no difference in question type for implants. Misconceptions were approximately 75% incorrect responses irrespective of topic or question type.

In discussing these results, more misconceptions in scenario-based questions than factual knowledge questions was expected. This was true of endodontic content, but not implant content. The absolute number of misconceptions was greater for implants than endodontic, but as a proportion of incorrect items, it was the same. It was concluded that the differences were related to numbers of incorrect responses, not confidence. More misconceptions were present for scenario-based questions because there were more incorrect responses with high confidence for both question types.

It was stated that an ideal situation is where a student’s ability is matched to their perception. Having found errors, three steps are required to help students reflect on these errors: identify misconceptions (incorrect and confident); provide feedback on their misconceptions; and motivate students to consider the exercise of self-assessment. It was presented that this as another means to assess self-monitoring. The self-reported limitations included the convenience sample of students and having two small assessments from a single institution.

3.7.2.2 Assessment including item-level certainty and safety

In 2012, Curtis and colleagues investigated misconceptions using certainty to responses. A misconception is an erroneous thought, idea or notion, resulting from misunderstanding or misinterpretation, and is difficult to detect in
assessments. Misconceptions can interfere with learning. High confidence incorrect responses reflect the presence of misconceptions. Traditional MCQ tests can be poor identifiers of misconceptions because: an uninformed student may guess correctly, and there is no feedback to prompt misconceptions; all distractors are considered equal; and, no feedback is related to confidence. Using confidence in responses can identify misconceptions. Further to this, distractors can represent various diagnoses and/or treatments leading to clinical outcomes, with potentially harmful results. Combining student confidence with a measure of clinical impact had not been attempted before. The context for this research was testing 105 third year dental students. Two staff categorised all distractors as benign (inconsequential or harmless to patient outcome), inappropriate (unsuitable or would delay appropriate care) or harmful (would result in direct or irreversible detrimental treatment). The agreement between the two staff was 52/60 (87%). Student responses included a sureness/confidence rating: very unsure (1), unsure (2), sure (3), very sure (4). In order to investigate if students were less sure for potentially more dangerous responses, an analysis of sureness for levels of impact of incorrect responses was undertaken. Misconception was defined as incorrect and sure or very sure (inappropriate or harmful).

Results were available for 105 students. The mean correct score was 90%. Of the 204 (of 2100) incorrect responses, 110 (53.9%) were misconceptions, comprising 8 (3.9%) harmful and very sure; 5 (2.5%) harmful and sure; 48 (23.5%) inappropriate and very sure; and 49 (24%) inappropriate and sure. Of the correct responses, mean sureness was 3.7; sureness for benign responses was 3.24, for inappropriate, 3.07, and harmful, 3.11. Sureness was higher for correct compared to incorrect, with no difference between categories of incorrect answers. There was no difference in sureness for levels of adverse impact. There was no difference in correctness, sureness, and misconceptions between male and female candidates. Seventy of the 105 students had one or more misconceptions, with seven having three or more.

In discussing their results, it was concluded that this method added additional information from sureness and clinical impact. The group with the lowest score was similar, but not identical, to those with the most misconceptions; the students were surer about correct responses; there was no difference in sureness between impact levels of incorrect responses; and no difference between genders.
One potential implication was that students with multiple misconceptions may be those who require careful review. These students were often incorrect and sure of potentially harmful decisions. Having misconceptions is different to being uninformed. Admitting to being unsure allows for a ‘teaching moment’.

Although the data were not presented in the results, there were four radiology-themed questions that the students did poorly on. It was postulated that these questions required higher-order cognition. Alternative explanations may include cohort-level misconceptions due to shared learning resources, or case-specificity demonstrated for misconceptions.

Further discussion is made of the lack of difference in sureness across levels of impact. The ideal expected would be sureness reducing with potential impact. It was postulated this lack of difference may be due to self-enhancement bias (seeing oneself in a positive light). Alternative explanations could relate to the categorisation into levels of impact or to the meaning of sureness understood by students. The lack of agreement of staff is discussed as a limitation in identifying potential misconceptions. Other limitations include a single assessment.

It was concluded that we seldom measure the potential for doing harm, and that there are educational benefits of using student confidence and clinical impact as outcomes in assessment.

Again, the analysis of sureness for correct/incorrect and also sureness across levels of impact could have been done differently, that is, that correct (or not) and levels of impact for levels of sureness when incorrect are alternative metrics from the same data.

In 2013, Curtis and colleagues highlighted the utility of certainty in- and consequence of- responses approach, and stated that this had not been previously attempted. In this study item-level certainty and safety was introduced as a way to improve self-monitoring. It was claimed that one of the advantages of certainty and correctness responses is that incorrect and more certain (misinformed) and incorrect and less certainty (uninformed) can be differentiated to improve educational opportunities. As misinformed beliefs are more resistant to change than uninformed beliefs, differentiating them could help direct further learning. Further framing of the potential purpose for this research included the student’s
identification of their own limits, and discouraging guessing. Testing and feedback can be directed at knowledge and self-monitoring. The analysis related to the importance of certainty to correctness, which is an important factor and will be discussed subsequently. Still within the framing, it was discussed that all distractors are not equally consequential, ranging from benign to harmful. A 20-item MCQ test including item-level certainty was given to 104 third year dental students as part of normal course requirements. The MCQs were single best answer from four options, and consisted of ten knowledge questions and ten complex vignette questions. The clinical impact was categorised by two staff independently marking each distractor as benign (inconsequential or harmless to patient outcome), inappropriate (either unsuitable or would delay the provision of appropriate patient care), or harmful (potential harmful results in direct and irreversible detrimental treatment of the patient). Kappa statistic for the 60 distractors was recorded as 0.76. It is not described what happened when the faculty members disagreed. Two is a low number of reviewers compared with the prior literature on categorising 'dangerous' answers, as discussed already.

A confidence scale was used: very sure (4), sure (3), unsure (2), very unsure (1). These were dichotomised to confident and not confident. Students were told that confidence was an important factor, but not told how it would influence their grades. It is not clear if the students knew in advance the scale would be dichotomised, and the fact that they were not informed about the influence of confidence on their grade is a possible driver to alter responding from honest to strategic. The purpose of the analysis was to differentiate uninformed from misinformed. Incorrect low confidence was deemed uninformed; incorrect confident was deemed misinformed.

Within the results, the reliability of 20 questions was Cronbach’s α 0.5 (though prior tests were quoted as 0.65-0.80) \( \Rightarrow \). Point bi-serial statistics were reported as previously being good. On the four-point confidence scale, the mean confidence was 3.37 for correct, and 3.08 for all incorrect, including 2.51 for incorrect benign, 3.02 for incorrect inappropriate, and 3.09 for incorrect harmful. Of all 2080 responses, 168 (8%) were uninformed (incorrect and unsure) and 450 (22%) were misinformed (incorrect and sure). Students were significantly more likely to be confident in a correct response than in an incorrect benign (odds ratio (OR) 15.9), incorrect inappropriate (OR 3.6), or incorrect harmful (OR 5.9) response.
Confidence did not decrease with potential harm. Confidence in incorrect benign responses (2.51) was lower than that for incorrect harmful responses (3.09), p<0.001.

In their discussion, Curtis and colleagues expected the result that confidence was higher in correct responses and simple questions. Gender had no influence, as found by others. That there were more misinformed than uninformed responses, and that confidence did not decrease as potential harm increased was found surprising. The distinction between the uninformed (incorrect not confident) and misinformed (incorrect confident) allowed for directed remediation. Altering the scoring function, with lower scores for incorrect confident, was postulated. The differences found for complex versus simple questions was not found by prior research, and still requires further exploration. Curtis and colleagues hypothesised why more harmful answers were not reduced, in that although confidence responding was specifically discussed, there was an assumption that students would implicitly consider the dangerousness of responses. Asking students to concurrently consider confidence and correctness allows students to calibrate appropriate levels of confidence in what they do and do not know.

Different reporting and analysis of data revealed that 5% of high confidence responses were potentially harmful (Table 3.5). Although the probability of incorrect responses was lower with confident responses, when incorrect confident responses did occur, inappropriate and harmful consequences were more likely (95% vs. 86%).
Table 3.5: Alternative reporting of data

<table>
<thead>
<tr>
<th>Of total responses</th>
<th>Not confident n = 290</th>
<th>Confident n=1790</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>122 (42%)</td>
<td>1340 (75%)</td>
</tr>
<tr>
<td>Incorrect</td>
<td>168 (58%)</td>
<td>450 (25%)</td>
</tr>
<tr>
<td>Benign</td>
<td>23 (8%)</td>
<td>22 (1%)</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>111 (38%)</td>
<td>347 (19%)</td>
</tr>
<tr>
<td>Harmful</td>
<td>34 (12%)</td>
<td>81 (5%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Of incorrect responses</th>
<th>Not confident n= 168</th>
<th>Confident n = 450</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>23 (14%)</td>
<td>22 (5%)</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>111 (66%)</td>
<td>347 (77%)</td>
</tr>
<tr>
<td>Harmful</td>
<td>34 (20%)</td>
<td>81 (18%)</td>
</tr>
</tbody>
</table>

In 2017, Rangal and colleagues explored item-level certainty and safety. This was within the framework of patient safety and self-monitoring of decisions, exploring conceptions and misconceptions. It is vital to identify the potential harmful clinical impact of misinformation held with high certainty. The setting for this work was with fourth year medical students sitting a neurology knowledge test. Two staff reviewers selected questions from a pool and classified the potential of each MCQ response as benign, risky, or harmful. Exam 1 comprised 30 questions and Exam 2, 30 questions (with one MCQ removed for analysis). Students could opt to sit the test under normal conditions (45 minutes) or certainty conditions (1 hour). Certainty was categorised as very unsure, unsure, sure or very sure (transformed to 1-4 for some analyses). Responses were classified by correctness and sureness (Figure 3.5).
<table>
<thead>
<tr>
<th>Incorrect answer</th>
<th>Correct answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sureness in correctness</td>
<td>Unsure of correctness</td>
</tr>
<tr>
<td>Usable</td>
<td>Unusable belief</td>
</tr>
<tr>
<td>Misinformed</td>
<td>Uninformed</td>
</tr>
</tbody>
</table>

Modified from *.`

*Figure 3.5: Classification by correctness and sureness*

For Exam 1, the analysis of 2700 answers revealed 2453 correct and 247 incorrect =: 1818 (67.3%) answers were informed; 635 (23.5) were guesses; 176 (6.5) uninformed; and 71 (2.6%) misinformed. Mean confidence was 2.98, with confidence in incorrect answers being significantly lower, at 2.02 than 3.07 for correct answers (p<0.001). When analysing by safeness of question topic, there was no significant difference in confidence when the answers were correct benign (mean confidence 3.14); risky (3.08); or harmful (3.05) (p=0.83). When analysing for incorrect answers, there was a difference in mean confidence: 1.8 for benign; 1.42 for risky; and 2.01 for harmful (p=0.06). Confidence in incorrect answers being correct was significantly higher for questions with harmful topics (p=0.01).

For Exam 2, the analysis, after one question was removed, of 2349 answers revealed 2066 correct and 283 incorrect =: 1453 (70.3%) answers were informed; 613 (26.1%) were guesses; 191 (8.1%) uninformed; and 92 (3.9%) misinformed. Mean confidence was 2.93, with confidence in incorrect answers being lower at 2.17 for incorrect answers compared to 3.02 for correct answers (p<0.001). When analysing by safeness of the question topic benign question mean confidence 3.12; 3.03 for risky; and 2.96 for harmful. Confidence in correct answers for benign questions was higher compared with risky and harmful. When analysing for incorrect answers, there was a difference in mean confidence: 1.78 for benign;
1.98 for risky; and 2.57 for harmful. Confidence in incorrect answers being correct was significantly higher for questions with harmful topic compared to benign.

In their discussion ..., other research was cited that has shown a difference in confidence by correct or incorrect answers. The hypothesis that confidence for incorrect answers would be reduced by clinical impact was found not to be the case. In fact, the opposite was found. The issue of student confidence being affected by many variables, and the measurements of confidence by further variables, was discussed including some as limitations: the effect of student personality, small magnitude of the effect sizes, effect of question type (harm), exam reliability, and student stress and motivation to score high, irrespective of clinical consequence. It was concluded by highlighting the benefit of re-enforcing lower confidence correct answers, and correcting incorrect answers, especially those held with high confidence.

In informing this thesis, it is worth noting that the stratification of harm was undertaken by two judges, and the means by which this was done was not described in detail. In addition, the questions rather than the responses were stratified. It seems likely that for any given question there would be variation in estimates of the potential harm of different responses. Single response safeness would be more informative than overall question safeness.

Having an increased likelihood of being correct when certain is equivalent to having an increased likelihood of: being correct when answering faster and/or not slowing down; being correct when not flagging questions; and being correct when responses are not changed. Having an increased likelihood of being certain when correct, is equivalent to an increased likelihood of: responding faster and/or not slowing down when correct; not flagging a question when correct; and not changing a response when correct.

The difference between being certain when correct and being correct when certain may link to the already highlighted issues of the spectrum of more efficient performance and safer performance (Figure 3.6).
Table 3.6: Self-monitoring for both efficiency and safety

<table>
<thead>
<tr>
<th></th>
<th>High certainty (or faster response and/or not slowing down, not flagging question, not changing initial answer)</th>
<th>Low certainty (or slowing down when responding, flagging questions, changing initial answer)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>60</td>
<td>10</td>
<td>70</td>
</tr>
<tr>
<td>Incorrect</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>

The probability of being correct when there is high certainty (safety) is $60/70 = 0.86$. The probability of having high certainty when correct (efficiency) is $60/70 = 0.86$. 

Modified from 81

**Figure 3.6: Interaction of sureness and correctness**

Examples of scores of how these different responses and analyses might occur are shown for three hypothetical candidates (Table 3.6, Table 3.7, Table 3.8)
The probability of being correct with high certainty (safety) is $60/90 = 0.67$.
The probability of having high certainty when correct (efficiency) is $60/70 = 0.86$.

These support that although candidates can all score the same number correct, the correct/certainty interaction can provide further information. All three candidates are of equal 'knowledge' ability, each getting $70/100$ correct on a 100 question test. However, their different response patterns with regard to certainty and

### Table 3.7: Self-monitoring for efficiency but less self-monitoring for safety

<table>
<thead>
<tr>
<th></th>
<th>High certainty (or faster response and/or not slowing down, not flagging question, not changing initial answer)</th>
<th>Low certainty (or slowing down when responding, flagging questions, changing initial answer)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>60</td>
<td>10</td>
<td>70</td>
</tr>
<tr>
<td>Incorrect</td>
<td>30</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>90</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 3.8: Self-monitoring for safety but less for self-monitoring efficiency

<table>
<thead>
<tr>
<th></th>
<th>High certainty (or faster response and/or not slowing down, not flagging question, not changing initial answer)</th>
<th>Low certainty (or slowing down when responding, flagging questions, changing initial answer)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>10</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>90</td>
<td>100</td>
</tr>
</tbody>
</table>

The probability of being correct with high certainty (safety) is $10/10 = 1.0$.
The probability of having high certainty when correct (efficiency) is $10/70 = 0.14$. 
correctness demonstrate how certainty when correct (self-monitoring for efficiency) and correct when certain (self-monitoring for safety) are different.

Despite the use of the concept of decision-certainty meta-judgements in clinical practice, there has been a paucity of information on how the data may be used.

3.7.3 Summary of assessment of self-monitoring

The vast majority of the research literature related to assessment and self-monitoring relates to using response analysis to infer the presence of self-monitoring rather than to measure it. For researchers using item-level certainty, both certainty-for-correctness and correctness-for-certainty are considered as being relevant for the analysis.
3.8 Summary of the chapter

In summary, with regard to this thesis, the main purposes of assessment are firstly, guiding and motivating learning, and secondly, making decisions about individual learner progress. Self-monitoring, which is framed within the concept of self-efficacy, can be informed by certainty in MCQ item responses and safeness of item response. Certainty-in and safeness-of item responses in a written/computer based assessment have been widely used in many different ways, with different formats and scoring systems. The importance of authenticity to clinical practice is highlighted for certainty judgements and certainty descriptors, and methods of assigning unsafeness classification. There is a small literature regarding assessment related to self-monitoring and some of this includes consideration of item certainty and item safety. However, these studies are limited, and there is a need to progress this field of work by considering inferences regarding self-monitoring that can be directly drawn from assessment including certainty-in and safeness-of responses.

This thesis will aim to fill some of the gaps related to using certainty-in and safeness-of MCQ item responses framed as being equivalent to the certainty-in and safeness-of clinical decision-making. These assessments may allow medical students to learn the importance of certainty and safeness, in a controlled environment whilst not directly impacting the care of real patients. The studies in the thesis will need to include analysis of responses at cohort and individual student level for: feasibility; scoring systems; progression decision-making; correctness for levels of certainty; safeness for levels of certainty; and the influence of student ability, experience, gender, ethnicity, and feedback on correctness-in and safeness-of responses for levels of certainty.
Chapter 4. Research paradigm and context
4.1 Chapter outline

This chapter will set the context for the research to be presented in two sections. The first section, 4.2, is a description of the research paradigm for this thesis, exploring the various paradigms that could be considered. Once the preferred paradigm is outlined, this will inform the research methodology and quality assurance that aligns with that paradigm. The second section, 4.3, describes the context of the medical school course to set the scene for the various assessments and students that are included in this thesis. This thesis consists of several studies, each of which has its own details with regard to methods used, and those will be presented in subsequent chapters. This chapter presents a general rationale for the use of a post-positivist approach, and describes the context in which the studies took place.
4.2 Research paradigm

The broad questions for this thesis relate to the utilisation of certainty-in and safeness-of MCQ responses as a means to introduce self-monitoring within medical student education. Subsequent specific questions vary as prior questions are addressed and considered. The optimal methods and analysis for any investigation will depend on the research, which in turn, will depend on the underlying research paradigm used. The need for paradigms to define the research process has been highlighted for discipline-specific research within educational research and for medical education research. There are various paradigms described in medical education research and for each there are assumptions about the nature of reality and knowledge that underpin the research.

- **Epistemology** refers to assumptions related to the nature of knowledge. Inquiry might relate to what is the: nature of knowledge; the relationship of knowledge to what would be known; the relationship of the researcher to what is known; and basis for truth? It establishes trustworthiness in data in context.

- **Ontology** refers to assumptions related to the nature of reality. Inquiry might relate to what is the: nature of existence; belief of the researcher; and way in which sense is made? Is there a reality out there or is it constructed by an individual’s mind? Is there an objective nature or individual’s cognition?

- **Axiology** refers to assumptions related to values-laden research. Inquiry might relate to what is the: the nature of ethics or ethical behaviour; values people live by, or are guided by, as research is conducted; nature of cultural and/or moral issues?

- **Rhetoric** refers to the assumptions related to the language of research. Inquiry might relate to what is the wording used in the data collection, results analysis and interpretation of results?

- **Methodology** refers to the assumptions related to the nature of research. Inquiry might relate to what is the design, approach, data gathering, participants, instruments, and analysis?

The choice of paradigm sets out the methodology, methods, and design. This education research is set in a complex and diverse field with multiple contextual
factors, and has lead towards adoption of a post-positivist paradigm. Post-positivism has tended to provide the world view of much research conducted on human behaviour in the education context. Humans involved in the social world accept that reality is imperfect, and truth is not absolute, but probable. For these reasons, a post-positive paradigm approach has been adopted for the studies in this thesis. Truth is also seen as something that sits outside individuals.

With regard to a post-positivist approach, the epistemological assumptions are that objective knowledge of world is not necessarily fully accessible and seeks to establish probable truth. Reality can never be fully understood. The ontology assumptions are that the world is ordered according to overarching truth. The axiological assumptions are that inquiries are value-free.

Post-positivist methodology develop knowledge through falsification of hypotheses. The emphasis is on well-defined concepts and variables, controlled conditions, precise instrumentation, and empirical testing. With regard to the methods of the research, although predominately quantitative methods are used, both quantitative and qualitative methods can be used to systematically gather and analyse data from representative samples. Methods have included experimental, quasi-experimental, correlational, reductionism, theory verification, causal comparative, determination, normative, tests, scale, correlational, causal comparative, randomised controlled trials, and survey. The research paradigm also informs the quality assurance of the research. Within the post-positivist paradigm, this includes internal validity, external validity, reliability, objectivity, which can be divided for quantitative methods into reliability (test-retest, parallel, internal consistency, inter-rater) and validity (internal, external, construct, conclusion), and for qualitative methods into reliability (dependability, auditability) and ‘validity’ (credibility, trustworthiness, authenticity).

Within each substantial section of the experiments conducted for this thesis, the alignment to the post-positivist paradigm will be highlighted and discussed.
4.3 Ethics committee approvals and research consultation with Māori

The studies undertaken as part of this thesis were approved in accordance with University of Otago policy related to ethical consideration of research and teaching proposals involving human participants (Appendix C).

Study 1  University of Otago Human Ethics Committee reference number D09/20
Study 2  University of Otago Human Ethics Committee reference number D12/379
Study 3  University of Otago Human Ethics Committee reference number D13/120
Study 4  University of Otago Human Ethics Committee reference number D16/037

Research consultation with Māori was carried out for Study 3 and considered by Ngāi Tahu Research Consultation Committee on 21 May 2013 with recommendations and suggestions provided (Appendix C).
4.4  Otago Medical School (OMS) MBChB degree Course (relevant years 2008-2015)

4.4.1  Course overview

From 2008 to 2015, the MBChB degree course at the University of Otago at the time of these studies and currently is a six-year course that is divided into four distinct phases. The first is a common Health Science year at the University of Otago Campus in Dunedin. This is followed by Early Learning in Medicine (ELM) years two and three which, apart from community contact weeks, is based at the Dunedin campus. Following this, students undertake years four and five, in the Advanced Learning in Medicine (ALM) programme at one of three geographically separate campuses in Dunedin, Christchurch and Wellington. The final year of ALM is a Trainee Intern Year, with students nominally based in one of these three campuses but in reality, spread across many different healthcare locations, working as members of healthcare teams.

The course components present 2008 to 2015, some of which are still present, are now described in greater detail.

4.4.2  ELM purpose

The purpose of the ELM programme was to build on the students’ prior learning and develop, for a further two years, the scientific foundations of medical practice and to begin the acquisition of the knowledge, skills and attitudes required for professional practice.

Specifically ELM aimed to:

1. Lay the foundations of basic sciences on which medical practice resides/relied and facilitate this learning by explicitly identifying the clinical relevance and contexts in which this knowledge will be applied safely in clinical practice.

2. Introduce clinical skills to give context and relevance to the medical sciences and to enable the application of medical sciences to clinical practice.

3. Establish the foundations of understanding the broad social and cultural concepts and contexts of medical practice and systems of healthcare.
4. Establish an appreciation of the patient as a unique person and partner in healthcare.

5. Begin the students’ professional identity formation and understanding of professionalism.

6. Establish habits of life-long learning, including both independent self-directed learning and collaborative learning.

7. Introduce the knowledge and skills of critical thinking and reflection required for clinical practice.

4.4.3 ELM overview

The ELM curriculum included three programme modules primarily delivered through small group tutorials: Clinical Skills, Early Professional Experience and Integrated Cases. Medical Sciences, including anatomy, physiology, pathology, microbiology, biochemistry, pharmacology and behavioural science, are taught via lectures, labs and small group tutorials. Students worked in tutorial groups of around ten students led by Teaching Fellows/Professional Practice Fellows and part-time tutors on the staff of the School of Biomedical Sciences and Dunedin School of Medicine.

The course components included:

- Clinical Skills tutorials involving instruction in basic clinical skills and doctor-patient skills using the students themselves, actors and volunteer patients from the community for medical history-taking and limited physical examination.

- Integrated Cases tutorials combining tutor-led, independent group-led and independent learning for each topic.

- Early Professional Experience (EPE) encompassing topics to help students understand the role and responsibilities of the doctor in practice. It included students spending time in a variety of the healthcare environments that they may experience as doctors, and meeting patients.

- Medical Sciences learning taking place in block modules based around body systems. The course is designed so each block module of lectures, labs and tutorials intersects with the cases, clinical skills and EPE learning.
• Vertical modules spanning throughout the two years provide ongoing learning in important areas of medical practice such as professional development, ethics and evidence-based medicine.

All tutorials emphasised an active small group collaborative working environment: this occupied about 15% of student learning time. ELM students spent just under half their learning time in formal teaching situations (lectures, labs), with the remaining 40% of time allocated for independent (non-contact) learning. The aim was to model future working practice that balanced working in teams with the ability to self-manage learning.

From early in the course, students had contact with patients, on campus and in the patient communities, around Dunedin and during Community Contact Week, around the country.

4.4.4 ALM purpose

The purpose of the ALM programme was to build on the students’ prior learning in Early Learning in Medicine and develop, for a further three years, the knowledge, skills and attitudes required for professional practice.

ALM’s overarching aims were to prepare graduates to undertake initial clinical practice as a medical practitioner under supervision and further study in any branch of medicine.

More specifically, ALM prepared students to:

• Further develop their knowledge of basic sciences relevant to clinical medicine and integrate this with clinical practice.
• Practise high quality, patient-centred/whānau-centred, evidence-based care of people experiencing common or important presentations and conditions.
• Recognise the patient as a unique person and partner in healthcare.
• Demonstrate cultural safety and competence within learning environments.
• Work effectively in healthcare teams.
• Develop a professional identity and understanding of professionalism.
• Further develop habits of life-long self-motivated learning.
• Employ critical thinking and reflection to improve their own, and others’, practice.
• Apply understanding of how the determinants of health can be incorporated within medical practice and systems of healthcare.
• Demonstrate commitment to equitable health outcomes.
• Demonstrate stewardship of medical resources and their appropriate utilisation.

4.4.5 ALM overview

Over the ALM years, students learnt through clinical exposure to patients across a wide range of sites and medical disciplines. Groups of students rotated through specific context-based block modules in hospital, community, in-patient and out-patient settings, both urban and rural. Students may also have applied to spend their fifth year in the Rural Medical Immersion Programme (RMIP), which entailed being part of a small group based in one of seven rural centres.

Students spent about three-quarters of their time gaining clinical experience, with the rest of the time spent in supplementary lectures and tutorials, including vertical module topics such as ethics, microbiology and pathology.

4.4.5.1 Module assessment: Year 5 General Medicine and Subspecialties module, Wellington

The ALM course was made of modules which include assessments that could be used for investigation of certainty-in and safeness-of responses.

The fifth year at one of the campuses included a five-week module in General Medicine and Subspecialties (respiratory, cardiology, oncology, nephrology). Six groups of students rotated through this module each year with approximately fourteen students per rotation. The module included clinical education in a variety of contexts including wards, out-patients, procedure clinics, the emergency department and the medical assessment unit. In addition, there was a tutorial programme to complement the learning-in-context. Assessment for the module included a MCQ examination, an observed structured clinical examination, case reports, peer review of performance, engagement in a variety of specified learning opportunities, and observations of professional conduct.
4.4.6 Whole class assessments

4.4.6.1 Retained Knowledge Test (RKT)

The main purpose of the RKT was to allow students to monitor and reflect on the knowledge as they progress through the MBChB course. RKTs were simultaneously administered to all year two to year five students twice a year throughout the course. Each RKT comprised MCQs covering content appropriate to all levels of the course. The tests were delivered over a two-week period via a learning management system (LMS), MedMoodle, to allow students to fit it in around other commitments. The tests were not administered under examination conditions.

The score by an individual student was not used to penalise a lack of ability. However, a student whose number correct is at, or less than chance, or who spent little time on the test (less than a mean 20 seconds per question) was sent to the relevant Student Progress Committee. This information was to be used, with other information, to inform consideration of their degree of engagement in personal education.

The questions for the RKT were taken from the OMS MBChB degree course assessment item database. Each MCQ incorporated a range of 5-16 options. The content related to any aspect that occurs as part of the core curriculum anywhere from Health Sciences Year 1 through to the completion of the course.

4.4.6.2 End-of-year assessments ELM2 and ELM3

At the end of each of the ELM years the following assessments were used to inform progression decisions:

- A written exam which consisted of:
  - Three 3-hour case-based SAQ papers, which may cover any of the learning requirements for relevant year, i.e., material from block and vertical modules, Integrated Cases, Clinical Skills and EPE.
  - One 1-hour OSPE (Objective Structured Practical Examination), which consists of approximately 50 questions on anatomy (including histology) and pathology.
• One OSCE (Observed Structured Clinical Examination) consisting of a maximum of eight 7-minute stations, drawn from the defined tasks outlined in the Clinical Skills objectives and will include consultation examination and consultation interview skills.

Each SAQ paper included several clinical cases, each part-question within a case was marked on a 1 – 5 scale (1 = clear fail; 2 = bare fail; 3 = bare pass; 4 = clear pass; 5 = potential distinction). Each question was weighted for aggregation, with the weighting calculated in accordance to the expected time to complete the question to a satisfactory standard.

The OSPE consisted of questions weighted equally and a candidate’s overall percentage result was calculated across the OSPE as a whole. The following conversion scale was used to convert the OSPE percentage mark into a score on the same 1-5 scale as used to mark the SAQ component: 100-80% = 5; 79-60% = 4; 59-50% = 3; 49-40% = 2; <40% = 1.

A sequential testing approach was taken in the OSCE. All students will initially attempt four OSCE stations, marked on a 1 to 5 scale (1 = clear fail; 2 = bare fail; 3 = bare pass; 4 = clear pass; 5 = potential distinction). Those students whose performance could be categorised as clearly and safely above the required standard from these first four stations were exempt further assessment in the remaining four stations. Students whose performance could not be categorised as clearly and safely above the required standard were required to undergo further assessment in the remaining four stations, with decisions made on performance in all eight stations.

4.4.6.3 End of Year assessments ALM 5

As for ELM, the end-of-year assessments in ALM5 comprised a written examination and an OSCE.

The written examination consisted of three papers each of three hours duration: two comprised 110 computer-delivered MCQs and one comprised five SAQs.

The broad aims of the written examination were to assess knowledge of: disease processes; the natural history of core conditions; diagnosis of core conditions relating to core presentations; management (investigations and treatments) of core
conditions and core clinical presentations; principles and practice of preventive and social medicine/public health; ethics; and the processes underpinning core conditions and core clinical presentations; and the processes underpinning clinical practice. Core conditions and core presentations were defined in the OMS MBChB degree course Curriculum Map.

The domains covered in this examination included: Clinical Skills; Professional Practice; Hauora Māori; Population Health; Research & Literacy; Medical Science; Investigations & Interpretations; and Management & Therapeutics. The broad discipline subject area categories covered in this examination were: Pathology; Internal Medicine; Clinical Pharmacology; Surgery; Psychological Medicine; Paediatrics; General Practice; Obstetrics and Gynaecology; Public Health; Bioethics; Hauora Māori; and Foundation Sciences.

Of the 110 MCQs in each of two papers, 100 questions counted towards the pass/fail and pass/distinction decision-making and the other ten served other purposes such as benchmarking or the trialling of content, format, longitudinal comparisons and standard setting. The 200 MCQs that were used to inform decision-making were determined in advance. The SAQ paper comprised five questions of equal value. Where there were separate sections within a question, the value for each section was indicated as the percentage of the total mark for that question.

The OSCE was designed to test students’ clinical skills. The clinical scenarios were chosen to represent a range of presentations seen in clinical practice, and were drawn from a range of disciplines. Three core competencies (history taking, clinical examination, and explanation and planning) were assessed every year. Additional competencies (e.g., Hauora Māori consultation skills) may also have been examined in any year. The examination was delivered simultaneously at the three main ALM campuses, and was identical in structure, stations and timing across the three campuses. The OSCE comprised eight stations, each of eight minutes duration. OSCE station content was derived from any of the core presentations listed in the OMS MBChB degree course Curriculum Map. Students were expected to respond appropriately to the ethnic/cultural background of the patient in all stations.
4.5 Summary for chapter

The research was based within a post-positivist paradigm. In subsequent chapters considering this paradigm will provide structure to the methodology, studies, quality assurance, and framework to interpret the results. The course description also provides context to the methodology, studies, and quality assurance to help interpret the results. The course context covers both the students’ programme, and where the students were located at the time of the studies. Where study results are compared to those of the regular course, context is provided.
Chapter 5. Developing assessment for certainty-in and safeness-of responses
5.1 Chapter outline

In Chapter 3 it was established that making diagnostic and therapeutic decisions is one role, if not the most important role, of clinicians as healthcare professionals. Clinicians require not only sufficient knowledge and skill to make decisions, but also the capacity to know when they have reached their limit: this self-monitoring is the activity of reflection-in-action as part of decision-making in daily practice. Belief about capabilities to undertake actions and belief about consequences of actions are at the core of the theories that underpin self-monitoring. Assessment of self-monitoring could be informed by certainty in item responses and safeness of item responses. Having established the need for an assessment that includes certainty-in, and safeness-of, responses as part of a framework of self-monitoring, this chapter will begin by briefly covering possible formats of assessment.

Chapter 4 described the context and research paradigm for this thesis, indicating how and where assessments could be possible within the course. After identifying a preferred format of multiple-choice question (MCQ), the chapter will explore how MCQs could be utilised, including structure, development and quality assurance.
5.2 Formats of assessment

There are many different formats and methods of assessment of medical undergraduates. These include written (which may include computer-delivered); practical/clinical performance; and workplace-based assessments. Written assessment include: TF (true-false) questions; MCQs; fill-in-gaps questions; short answer questions (SAQ); and essays. None of these is perfect for every purpose; each has benefits and limitations. Therefore, there is a rationale to use different methods to complement each other. The vast majority of both certainty-based and unsafe/dangerous-response assessments are based around MCQs. MCQs have been, and probably continue to be, the most prevalent form of 'objective' testing. There is significant experience and familiarity with this regular and widespread use.

Given the prevalence of relevant literature, and the widespread use of the MCQ format within the Otago MBChB degree course, consideration of the use of MCQ is pertinent for this thesis. Additionally, the use of MCQs as the format for investigation aligns with the post-positivist paradigm. Post-positivist methodology includes the development of knowledge through falsification of hypothesis. Therefore, the studies within this thesis need to be framed as such. With regard to the methods of the research used within the post-positivist paradigm, although predominately quantitative methods are used, both quantitative and qualitative methods can be used to systematically gather and analyse from representative samples. The study of MCQ responses lends itself to predominantly quantitative analysis, but with qualitative methods adding to the data analysis.

As already noted in section 4.2, the post-positivist paradigm assumes the need for demonstrating internal validity, external validity, reliability, and objectivity, which can be divided for quantitative methods into reliability (test-retest, parallel, internal consistency, inter-rater) and validity (internal, external, construct, conclusion), and for qualitative methods into reliability (dependability, auditability) and 'validity' (credibility, trustworthiness, authenticity). These align with the quality assurance methodologies considered for MCQ items and tests.
5.3 MCQs

MCQs involve selecting a response or responses from a provided list. Generally, the questions are made up of a stem (wording describing the topic), then a lead-in question, then a list of options for possible responses.

There have been several articles considering the advantages and disadvantages of the MCQ format. The advantages include factors related to construction and delivery, such as large areas of content covered, many cognitive abilities covered, automated marking and feedback, sample from broad content in time available, ease of marking, less expensive, and standardized delivery; and factors related to scoring and inferences such as reliability, analysis at the item level, and computer analysis of raw data allowing easier comparison between groups and questions. The disadvantages include factors related to construction and delivery.

In construction disadvantages include good items being difficult to write, time required to construct tests, and some limitation to assessment of only factual knowledge, thereby encouraging rote learning and cramming. In delivery these include factors related to scoring and inferences, such as guessing being encouraged, adverse impact of scoring schemes, and different question mix between candidates. Despite what some authors report, it is not the MCQ format that encourages guessing, but the scoring methods used.

In the context of this thesis and certainty-in and safeness-of responses, the MCQ format advantages, especially the increased content sampling and coverage per unit of assessment time, outweigh the disadvantages.

5.3.1 MCQ stems/topics

Stems should be clinically relevant and authentic. Paper-based scenarios run the risk of dehumanising patients, which will alter the HCP approach. Therefore, case scenarios need to maintain relevance and human features. Despite this, scenarios describing patients may need to be reduced because of size and complexity.

5.3.2 MCQ lead-ins

“Which of the following is the most … (e.g., likely diagnosis)?” is the preferable style for lead-in questions. This means that distractors do not have to be
categorically incorrect or unsafe, but lie on a spectrum of correctness, and for the current study, on a spectrum of unsafeness, this being more authentic to clinical practice.

5.3.3 MCQ response options

The number of options (and therefore distractors, incorrect options) is a source of ongoing debate. There are numerous articles comparing the number of response options and distractors. It has been postulated that reducing the number of options does not lead to a reduction in test quality parameters. This proposition builds on prevalent opinion as articulated in reviews and primary research reporting that reducing the number of options does not result in significant differences in test parameters, or that it can lead to improvement in the parameters. Although there may be issues with whether some of the research has sufficient power to detect differences, overall it seems to be the case that supporting reducing the number of options in an item.

While a low number of options seems appropriate, one concern regarding MCQs is that the correct answer is in front of the candidate. This may lead to cueing, which in turn will lead to a candidate choosing a correct answer that they may be unable to produce spontaneously, thereby reducing the validity of the scores. The fewer options that are available, the more likely that recognition of the answer by cueing may occur. Further, a lower number of options increases the chance of random guessing leading to a correct response.

The improved reliability found by reducing the number of options may be an artefact of a concurrent increase in numbers of questions. If response time is proportional to option numbers, non-functioning distractors that are rarely selected increase assessment time. Therefore, fewer options means more questions per testing hour, so greater reliability and content coverage.

It has been proposed that psychometric analysis will allow for reduction to an optimal number of distractors whilst maintaining item discrimination. Removing non-functioning distractors does not change psychometric characteristics of a test item. However, what is the goal of the assessment? The number of 'plausible' distractors will depend on the setting and purpose of the assessment.
Removing distractors assumes that rarely chosen distractors would not be important. If such selected distractors were part of a pattern of highly unsafe responses, this would be valuable information. There is more to the choice of number of distractors than psychometric analysis, and one relevant aspect would be clinical importance. Distractors identified by an expert panel as desirable by psychometric analysis and clinical importance can be different \textsuperscript{36} and personal communication from corresponding author (Professor David Swanson, Deputy Vice President for Test Development, National Board of Medical Examiners, Pennsylvania, USA, 12 May 2012). Recommendations for the exclusion of distractors include less than 5% of examinees choosing a distractor \textsuperscript{36}. Distractors that do not add to traditional psychometric analysis should be included. It would be important to note if the same 5% of candidates is choosing certain distractors that are unsafe. Options that are rarely chosen can represent unsafe practices: it is vital to know which students are selecting these potentially unsafe responses \textsuperscript{144,265,266}. Individual misconceptions can be included in feedback with the goal to direct personal learning development \textsuperscript{144,265,266}

The number of options in MCQs should be sufficient to include both psychometrically important distractors and clinically important (safe/unsafe) distractors. As the number of each will not be the same for all content areas, their inclusion will lead to variable option numbers. All items do not need to have the same number of distractors \textsuperscript{297}.

5.3.4 Extended matching MCQ

Extended Matching Questions (EMQs) were developed as a progression in MCQ format \textsuperscript{295,298–302} intended to align to assessment of forward, and to a lesser extent backward, clinical reasoning \textsuperscript{302}. EMQs were developed to assess clinical reasoning, but with an attempt to counter some of the problems with free-response questions \textsuperscript{298,300}. Some of the problems included: examinees need to estimate what the question author and markers intend regarding how much detail and specificity is required; scoring is resource intensive and subjective; and as questions are written to be more focused, less content coverage is possible. Although computer completion \textsuperscript{303} and marking \textsuperscript{304} are becoming possible, computer marking of free text answers is not widely used as yet \textsuperscript{305}. For these practical reasons, EMQs are preferred to free text \textsuperscript{306} response formats.
The structure of EMQs includes a theme, a set of options, a lead-in, and stems. To create authenticity and familiarity, the theme is a general topic - usually a presenting complaint - which can be based on a patient problem list. Construction of EMQs from lists of common clinical problems and corresponding diseases is possible. The option list generally comprises more than six options, with up to 20 or more, or up to 26 reported. There is no limit to the number of times an option is used as the correct response nor for the use of stems for different option lists. The lead-in statement is the task, or question in the one-best-answer format, and provides direction. A stem is usually a clinical vignette, although non-clinical scenarios can be used. Clinical vignette interpretation requires analysis and synthesis, and the avoidance of irrelevant information in the vignette is recommended for construction.

The proposed advantages of EMQs include: less cueing, less chance of guessing correctly, reduced impact of guessing, more reliable results, more discriminating results, assessing pattern recognition, assessing application of knowledge, assessing problem solving, and focusing student attention for learning. Time resource is important in EMQs, and long scenario stems, such as those used in most EMQs, require more time to read in the exam, with 90 seconds per question being recommended.

5.3.5 Long menu options

Some of the benefits of EMQs relate to the number of response options. Long list options have been developed which can be linked to the theme of the stem. This could result in variable numbers in the option lists. A long menu of options might be possible, covering all MCQs. Such long menu MCQs have proved feasible when facilitated by computer delivery although they were even possible in a paper-based system.

5.3.6 Order of presentation of stem, certainty judgement and options

The certainty judgement can be made at different times within the flow of stem, lead-in question, options review, and option choice. An example is: stem, certainty judgement, lead-in question, options to review, option choice. This is asking, "How certain are you in this topic?". The question and options are not known at the time of the certainty judgement. The advantages are that candidates must
think and formulate first, with no clues before certainty judgement, and cannot work backwards from cued answers.

Another example is: stem, lead-in question, options, option choice, certainty judgement. This is asking, "How certain are you in your response?". Selecting the option choice before the certainty self-judgement improves predictive accuracy. The response choice provides further information in making the certainty judgement.

In linking self-monitoring to self-efficacy, as previously discussed, the decision of the candidate responding to an item is whether the limit of their ability has been reached. This comes after the initial decision and consideration of certainty and safety. Hence the stem, lead-in question, options, choice, certainty judgement order is most appropriate.

5.3.7 MCQ scoring

To be meaningful, certainty-in and safeness-of responses need to be built into the scoring system for the assessment. Number correct scoring and formula scoring (correction for guessing) are the systems most widely used.

5.3.7.1 Number correct/right scoring

Each correct response scores +1 and each incorrect (or 'don’t know' option, if present) scores 0. This is a simple and easily understood scoring system.

There are problems reported with this, however. Test scores based solely on the number of correct answers ignores potentially valuable information, in particular, the quality of the incorrect responses. If the possible incorrect answers are not equal, information will be lost. Usually all incorrect answers are given equal weight if chosen. An incorrect answer may be a guess, incorrect information held with variable certainty, or a typing error.

The all-or-nothing scoring system is imperfect, as it forces a guess. The continuous variable of various degrees of knowledge are not recognised by dichotomous scoring. Likewise, real life is not always an all-or-nothing situation. The distinction that answers are known or not is artificial and leads to false assumptions regarding the respondent’s level of knowledge. Partial knowledge
will increase the chance of being correct 20, but there is no recognition of partial knowledge 152,164,170,175,180,186,188,212. It is likely that in an achievement test there will be some partial knowledge demonstrated by the candidates, even though each question has a uniquely correct answer 152. It has been argued that partial knowledge and its effect on scoring should be recognised 152,164,170,175,180. Responses in MCQs based on partial knowledge and actions based on partial knowledge are different 188.

5.3.7.2 Formula scoring

Number correct scoring can be seen as an invitation to guess 20. Concerns over guessing led to the development of formula scoring 20, whereby an incorrect selection would receive a score (usually negative) rather than zero, as in number correct scoring. As the MCQ format became more widely used it was criticised because guessing could increase the score 20. As a different approach to scoring, formula scoring may be justified, as admitting not knowing is preferable to guessing 20.

Formula scoring was not designed to penalise guessing, but to adjust scores for gains due to random guessing 20, although this has been debated 20. Even though not designed as penalty system, scoring systems to discourage guessing have been divided into no-penalty and penalty systems 20. Some formula scoring is also known as 'correction for guessing' (CFG) scoring, but as noted, it does not correct for guessing. At best, CFG discourages guessing 20 by some candidates to a variable degree 20. Many MCQs used in medical post-graduate examinations have employed formula scoring 20, although this is changing towards number correct scoring internationally 20.

Formula scoring was developed to improve the statistical reliability of scores 20. The formula subtracted a countermark of variable size for incorrect answers from the total correct score. The size of the countermark varies 20. Although essentially equitable mathematically, the following are also used: mark for correct of 1, countermark for incorrect of 1 divided by the number of options minus 1, score for 'don’t know' option (DKO) of 0; and versions of this, which remove the negative and give credit for DKO response, mark for correct of 2, countermark for incorrect of 0, score for DKO of 1. Others used different penalties such as (for number of
options equal 2): mark 1, countermark $\frac{1}{2}$ or $\frac{1}{3}$. Whatever formulas are used, examinees should be informed and understand them.

Some scoring systems are based on the assumption that all distractors are likely to be chosen equally, given a lack of knowledge. However, some distractors may be chosen more frequently than others. Some distractors will be more clinically important than others, potentially justifying different countermarks for incorrect response options.

5.3.8 ‘Don’t know’ option

Candidates ask, “How do I proceed if I don’t know the correct response?”

Having a DKO would allow candidates to be honest, and prevent candidates from being forced to answer as if they have complete certainty in a particular answer. On the other hand, the removal of the DKO is justified, as it removes the personal tendency to use DKO, which can cause irrelevant variance in test results.

There is advantage to candidates answering every question, even where they are completely ignorant of the answer. Even so, some cautious candidates will leave answers blank, and, as a result, are potentially penalised compared with their less cautious peers. However, the requirement to answer all questions in the face of ignorance should not be part of medical education.

Choosing the DKO should be recognised as being inferior to selecting the correct option, but superior to selecting the incorrect option. This is one argument for formula scoring. However, formula scoring was designed to adjust scores for gains due to random guessing.

5.3.9 Guessing and scoring

Guessing is not encouraged for two reasons. The first reason is that guessing, when uncertain, is not a behaviour authentic to good HCP practice. The second and more prominent reason reported in the literature, is that guessing will reduce the reliability and validity of the item and test. Guessing is not related to ability, but is associated with personality characteristics.

Guessing will increase a score from a given true score. Score accumulation through guessing, encourages guessing. Guessing may be random (no
certainty) or based on partial knowledge or a degree of certainty\textsuperscript{41,19,20,28}, or alternatively represent misunderstanding, bad habits, or carelessness\textsuperscript{10}. As risk-taking is encouraged in society, students will guess in assessments, but is it desirable for healthcare professional students to guess when unsure about a patient’s care\textsuperscript{30}?

Informed guessing can lead to small improvements over random guessing. In measures of learning\textsuperscript{34}, and also improve test score reliability\textsuperscript{31}. Random guessing reduces the reliability of scores\textsuperscript{32}. Scores can be increased by guessing\textsuperscript{33}, and there are a number of factors to consider when creating MCQs to mitigate issues, including the number of questions, number of choices, and the scoring system. Score probability distributions can be calculated for given levels of ability, and number of question options\textsuperscript{34}. Formula score and number correct are perfectly correlated where there are no omitted question responses\textsuperscript{35}. Formula scoring will not change the number of false positive candidates incorrectly scoring above the pass threshold, so staff and the public have misplaced overreliance on benefits of formula scoring\textsuperscript{36}.

These issues point to the difficulty in commonly used scoring systems, including number correct and formula scoring. The over-reliance on score reliability is problematic, when it is increased by reducing the diversity of questions at the expense of content representation, and manipulation of raw scores, with potential consequences in terms of response patterns.

5.3.10 Strategic responding, test-wiseness and personality

Although making decisions based on unreliable scores is problematic, increasing score reliability does not guarantee better decision consistency\textsuperscript{30,31}. Procedures that serve only to improve score reliability might do so with no benefit in terms of decision consistency, and might actually have unforeseen consequences with increasing irrelevant variance.

Scoring systems encourage strategies related to the system and not to clinical practice\textsuperscript{211}. Test-wiseness is the use of skills and strategies, not related to the construct being assessed, to enhance the test score\textsuperscript{40,188,206,297}. Distinguishing correct answers based on knowledge from those based on guessing is not possible\textsuperscript{30},
likewise distinguishing correct answers based on knowledge from those based on test-wiseness will not be possible. One assumption about number-correct scoring is that candidates will answer every question 
, but this is not always the case 
. Some will omit answers, based on willingness to answer or test-wiseness, thus potentially reducing the validity of the results.

Attributes other than the application of knowledge can increase scores. There is a misconception that formula scoring leads to the same results as refusing to guess 
. Guessing still affects scores under different 'do not guess' instructions 
. There is variable advice to candidates about formula scoring responses: never guess; play hunches; or answer all 
. The perception exists that candidates who do not understand the concept of formula marking are significantly hindered in attaining their best score 
. Instruction wording is important 
. For example, instructions making clear that “the penalty for wrong answers is severe” as opposed to “on average there is no benefit from blind guessing over omitting answers” will be likely to prompt different response patterns. Formula scoring instructions are interpreted differently and alter responses. Instructions could advise a candidate to answer if they have valid partial information 
. Individuals have different thresholds for selecting a DKO. The justification for answering all is that even when formula scoring with a balance penalty is used, candidates are still more likely to be correct on questions that were perceived to be a random guess, both in best-of-five format MCQs with 54% unanswered correct 
, and TF format, with 65% correct 
.

Despite false answers being easier to write for TF than MCQs 
, there can be an excess of true answers in true-false questions, so strategic responding is to guess true for all 
. A question and scoring system needs to minimise the impact of strategic responding. Such advice focuses candidates’ concentration on the scoring scheme rather than the content of the examination 
. The examination becomes a game, where the candidate is trying to score as highly as possible, and the examiner is trying to ensure that their score reflects knowledge 
.

Candidates will spend time considering a scoring system rather than demonstrating knowledge 
, which can be compounded by some systems requiring an increased time to complete 
, thus penalising the cautious 
. The obsessional candidate may procrastinate even when their degree of certainty
is high \(^\text{152}\). The anxiety trait does not influence scoring on MCQs, but a negative association between score and anxiety has been reported for female candidates assessed on tests that use formula scoring \(^\text{152}\). Candidates believe formula scoring is a penalty applied to lower their mark \(^\text{211}\) and is not motivating or rational to them \(^\text{191,228}\).

Although the scoring function can alter response patterns by candidates, it is the candidates’ perception of scoring function instructions, rather than scoring function per se. It has been demonstrated in a randomised controlled trial that instructions lead to a change in response pattern, even when scoring was not different \(^\text{144}\).

5.3.11 Making decisions in clinical practice and making decisions to respond to MCQs

Clinical decision-making requires decisions and therefore 'don’t know and move on' is not an option \(^\text{277}\). This has been explored further \(^\text{277,326}\), and the assumption that examination candidates will know an answer or not, and if not, that any answer is a guess, is flawed, and disregards partial knowledge. This point is endorsed by others \(^\text{111}\).

A good clinician may take a 'calculated' risk on treatments with a chance of success for a high morbidity/mortality disease. In the same way, medical students should be discouraged from random guessing, but not inhibited from acting on incomplete information \(^\text{322}\).

Better ways to make allowances for 'guessing' \(^\text{152}\) are needed, and these ways should reflect the reason why random guessing is not to be encouraged, that is, their appropriateness in HCP practice. It is wrong to encourage random guessing, and it is important that a HCPs can recognise and admit when they do not know an answer \(^\text{277}\). Random guessing is an unsound basis for clinical practice and should not be encouraged under any circumstances \(^\text{5,311}\).
5.4 Writing quality MCQs

There have been several articles regarding constructing quality MCQs, what makes a quality MCQ and avoiding flaws whilst constructing and developing MCQs. These cover very similar themes, and three have been selected to cover the range of issues, along with a plan for good practice within the current study (Appendix D).\textsuperscript{44, 125, 282.}

From these the items and tests to be used in this thesis were developed such that:

- Item content will be taken from curriculum map and/or learning outcomes, and verified by clinicians
- Formats will be pen/paper or computer delivery where feasible
- Items will be selected to ensure no items link to each other
- Lead-ins will avoid: vague semi-quantitative terms; words such as ‘except’; and negative words, but include negative concepts such as side-effects and contra-indications
- Lead-ins will be written such that the stem provides vital information to get correct answer, it should not be possible to look at lead-in and options, and not the stem, and get correct answer;
- Options will be written to avoid: grammatical, logical, convergent cues; and ‘all’ or ‘none’ response types
- Options will be written with one 'best' answer, verified by clinicians
- Clinicians will verify the authenticity and wording of the stems, lead-ins and options
- Options will be placed in alphabetical or numerical order, to make responses easy to find in long lists.
5.5 Standard setting

5.5.1 Standard setting for knowledge

For the assessments to be used in this thesis, standards and standard setting for the correctness, certainty, safety interactions and self-monitoring need to be considered.

There are a very large number of potential standard-setting methods. The Modified Angoff score is used to set a standard based on the judgements of a panel, and it is used for several MCQ knowledge assessments within OMS as described in section 4.4.6.3 Therefore, the staff are familiar with the process and the concepts and its use within the context of this thesis.

In the most widely used version of Angoff at OMS “panellists review multiple-choice items and provide an estimate of the probability of minimally-competent candidates answering an item correctly”. Minimally-competent is used as the threshold performance, rather than borderline, to describe the candidates.

The panellists are chosen for being familiar with content, contexts and also the standards. Therefore, the panel needs to have clinical expertise and also expertise in the practice of medical students and early postgraduate doctors in New Zealand.

It is important to demonstrate that panel composition is considered and is justifiable. The composition of the panel adds to the credibility of the derived pass threshold. Although training and a common understanding is achieved, panellists will have idiosyncrasies: some variance between panellists will still occur and is to be expected. Therefore, having a sufficient number of panellists is important. The number of panellists and the time provided for their deliberations is a resource constraint. Effectively the number of panellists is ‘as many as can be afforded’. Angoff methods can include providing judges with the responses of other panel members, followed by questions undergoing a second review, although this can be time consuming.

Angoff standard-setting does not take question content relevance into account. Usually content is verified in advance, but can be confirmed by panellists. Panellists can be utilised as question reviewers, which is important given the
need for clinical experts to review questions for authenticity, and given the limited pool of people to undertake such tasks.

Knowledge of prior performance data for questions can help\(^{155}\), but this will not be possible, as the study reported in this thesis is effectively the first use of MCQs to include certainty-in and safeness-of responses.

5.5.2 Standard setting certainty, safeness, interactions with correctness and self-monitoring

The only reported development of standard-setting related to certainty based assessment (CBA) was associated with logarithmic scoring\(^{148,226}\). The pass-mark was set as per number-correct scoring, but scores are then adjusted to equate that standard with the CBA score. This power log scaling occurred over the range of scores\(^{148,226}\). Even if mathematically sound, it does raise the question of the validity of combining correctness (knowledge) and correctness-certainty interactions (one part of self-monitoring) into a single score that can be standard-set.

There was no identified literature of standard setting with regard to safeness of responses. As covered in sections 3.6 and 3.7.2.2, there are reports of retrospective and prospective identification of unsafe responses, none of these described a standard or standard setting.

There was no identified literature of standard setting with regard to self-monitoring. As covered in sections 3.7.1 and 3.7.2, there are reports of inferred evidence for self-monitoring and using certainty-in and safeness-of responses, but none of these described a standard or standard setting.
5.6 Feedback to guide learning

Feedback is a significant part of assessment and is one of the most effective stratagems to guide future learning. Depending on the model, feedback with provision of correct answers to any incorrect responses may have no effect on learning. Feedback in an open system concentrates on reinforcing correct responses, whilst a closed system concentrates on correcting errors.

The aim of feedback can include highlighting incorrect responses for future learning or affirming correct responses, and depending on which of these is the primary aim the timing of feedback is important. In the case of providing feedback to correct errors, a delay in giving feedback may be optimal. This would differ if the aim of feedback is to provide positive reinforcement. Feedback on incorrect unsafe responses, with higher certainty, is potentially more important than affirming correct responses, even if only associated with lower certainty. Safety is generally more important than efficiency.

A lack of specific feedback between tests does not lead to any increase on scores on identical items, and if anything, students perform worse. Producing feedback specific for the person is advantageous, with the target to improve assessment performance, which would infer improvement for practice.

As already discussed in section 3.2.1.2, feedback is information on the assessment content that gives learners information, on gaps or not, in their progress of study and associated clinical work. In this setting of clinical decision-making and self-monitoring, feedback should include information on more than just content correct/incorrect, but also certainty-in and safeness-of responses in order to be used to guide future learning. Components of feedback include verification and elaboration. Verification is a simple correct or incorrect. Elaboration provides further substantive information, and in this case, certainty meta-judgements and indications on safeness. This provision requires consideration of: the type of information, specific to the task/item; the form of information, the presentation of question; and the load of information, how much to provide. However, more feedback information more does not always lead to better follow-up test results.

Giving feedback to students on prior response certainty will allow students to focus on future learning activities, and can influence students' future responses
The greatest potential for student learning is for incorrect responses held with high certainty. Feedback could focus on both future student learning and improving student meta-cognition. It has been recommended that for feedback to guide learning, the feedback should better directed by being based on the appropriateness of certainty in responses in the assessment. Students devote most time to consider high-certainty incorrect responses, with less taken on correct answers, whether high or low certainty, and items that are incorrect but held with low-certainty being in between these in terms of time allocated. Discrepancy is the error in the meta-judgement, misplaced certainty in knowledge. When knowledge and certainty are high, discrepancy is low, and feedback and further study time can be focused elsewhere, improving learning efficiency. For high-certainty error, high discrepancy, more elaboration, in terms of detailed feedback, is required. Mid-range discrepancy may have the greatest chance of correction. Memorability and certainty are related, so breaking these links when misplaced can be difficult.

Useful feedback could lead to a shift from decisions that are incorrect and unsafe with high certainty to those that are incorrect not unsafe with low certainty, and from responses that are correct with low certainty to correct with high certainty. Feedback should allow students to assess their own insightfulness, decreasing misplaced certainty, even if errors are not corrected. On the occasions when they make mistakes with high certainty, it is instructive for candidates to consider why they failed to see the risks they were taking. This categorisation of misconceptions about the safety-certainty levels of responses is useful. Identification of reasons underlying misconceptions can help learning through provision of targeted feedback. Feedback should allow for the correction of errors, with the seriousness of errors given importance and made explicit. Errors can be stratified by seriousness, which may be straightforward for non-complex mathematics questions. Non-serious errors, however, can be improved by provision of comprehensive feedback, but serious errors on the other hand may be relatively unaffected by feedback.

Given the importance of the certainty/safeness interaction, a matrix of the interaction of certainty, correctness and safeness of responses could be a good tool to use to help present the feedback. Quantifying degrees of error as an error.
metric is seen as unique. The tolerance of error and certainty/unsafety interaction in section 3.6.4 highlights this.

Feedback is not always successful in helping candidates to change their unsafe and incorrect knowledge. If the feedback is not remembered, or acted on, certainty in incorrect responses can continue, and even increase after feedback. Following feedback, self-assessment, as response accuracy, has been reported both to improve and not improve. Conversely, people may correct responses irrespective of feedback.

The students who most need to improve are not always those most willing to. Certainty will affect the effect of feedback. Certainty is a useful index of learning behavior as it reflects affective and motivational states in addition to knowledge. The effectiveness of feedback is influenced by degree of certainty in response. High-certainty errors can be less likely to be corrected. Durability, the likelihood that a response will be repeated at a future time, and certitude are positively related. This holds true for error and correct responses.

Although feedback is a significant contributor to the assessment purpose of guiding learning, including that related to self-monitoring, guiding learning can start before and continue during an assessment, that is, before feedback can occur. Self-monitoring itself may drive self-regulated learning. External feedback should relate to both the task and the self-monitoring related to that task. This feedback will guide ongoing self-monitoring.
5.7 Summary for chapter

For use with certainty-in and safeness-of responses, the advantages of the MCQ format outweigh the disadvantages. Care should be taken in MCQ construction: items should include clinically relevant authentic stems, followed by 'choose the best option' as a lead-in question, with no limit on the number of options, with a certainty judgement following the selection of a response. Widely used scoring systems such as number correct and formula scoring are not without problems. Standard-setting, although established for knowledge-related aspects of assessments, is limited for certainty, safety and self-monitoring, and will need to be developed. The systems chosen should be published to candidates. Feedback to candidates should be focused on their needs, but include information related to correctness, certainty and safeness, and the interactions between them. Highlighting higher-certainty unsafe responses is important. The response and scoring system chosen should be supported by psychometric analysis including group, subgroup and individual analysis.

The scene has been set for studies to explore the development of MCQ assessments incorporating certainty-in and safeness-of responses.
Chapter 6. Study 1
6.1 Introduction

Having set the scene for the need to explore the development of MCQ assessments incorporating certainty-in and safeness-of responses, a plan was made to undertake a series of studies. Over the course of these studies the practicality and feasibility of certainty-in and safeness-of responses, response and scoring systems, and psychometric analysis including group, subgroup and individual analysis would be required. Each study would be informed by theoretical and experimental work that had gone before and the studies within the thesis that had been undertaken.

This was the first study to explore a multiple-choice question (MCQ) response system incorporating certainty-in and safeness-of responses within a framework of self-monitoring of clinical-decision-making. The term ‘confidence’ was used in Study 1, with ‘certainty’ being used subsequently.

6.1.1 Hypotheses and research questions

The hypotheses for Study 1 became:

- Including confidence (certainty)-in, and safeness-of, responses in an MCQ assessment would be feasible.
- An item-scoring system for confidence (certainty)-in, correctness of, and safeness-of, responses would be feasible, and possible to aggregate in a meaningful way.
- The information provided by including correctness for levels of confidence (certainty) and safeness for levels of confidence (certainty) would be of greater value than the information provided by correctness of responses alone.

The research questions for Study 1 became:

1. How can confidence (certainty) and safeness be included in question construction and delivery?
2. Does including confidence (certainty)-in, and safeness-of, responses in assessment instructions alter students’ responses?
3. How can confidence (certainty)-in, and safeness-of, responses be incorporated into a scoring rubric?
4. Can we quantify students’ correctness of responses for levels of confidence (certainty)?
5. Can we quantify students’ safeness of responses for levels of confidence (certainty)?
6. How do these measures of correctness and safeness for levels of confidence (certainty) relate to students’ levels of experience?
7. How do these measures of correctness and safeness for levels of confidence (certainty) relate to students’ levels of knowledge?
6.2 Methods

6.2.1 Design

The design chosen for this study was a randomised crossover design, with students randomised to first having a group of questions in conventional format or questions including confidence (certainty) in, and safeness of, response format, and then, secondly, a different set of questions in the other format. This design was chosen as it allows all students to be exposed to both formats thereby enabling a comparison of format that is not confounded by the ability level of each student.

The context of this study and the MBChB degree course are described in section 4.3.

6.2.2 Question pool

6.2.2.1 Question pool content and review

To create a question bank for this study, 29 questions were taken from the OMS MBChB degree course assessment items database. These questions had already been developed and reviewed by OMS staff. All 29 were based on clinical scenarios and had content targeted towards the Year 5 common component examination. Questions were selected to reflect a broad range of topics and outcomes for the course, although there was no specific blueprinting with regard to questions selected. Subsequently, these questions were removed from use for high-stakes assessments within the course.

The questions were modified with the addition of a variable number of distractors (incorrect responses) such that all responses had ten possible responses with the tenth being the same for all questions: “I do not know the answer and would look it up or ask the appropriate person for help”. Each question therefore had one single best answer, eight distractors, and the 'don't know' option (DKO).

A panel of four hospital-based senior clinicians, who were also OMS staff members and practicing clinicians, reviewed question content and suggested changes. Following developments, the questions were agreed upon and correct answers confirmed.
6.2.2.2 Question standard-setting process

The standard-setting process used was the Modified Angoff. This process was used at that time for standard-setting for the OMS Year 5 common component examination, and the same four-member panel was used as the panellists, they being familiar with the process.

The standard to be used to produce judgements was “minimum to exit Year 5 and enter Year 6 of the course”. Discussion ensured that the panellists had a common understanding of the process and standard. Independently each panellist gave a judgement of the percentage of students with a performance standard of “minimum to exit Year 5 and enter Year 6 of the course” who would get the question correct. This was done using a Likert scale of 0-100%, with 10% increments. Mean judgements were used to create the estimated pass-mark for exiting Year 5.

6.2.2.3 Question classification for safety process

Simultaneously with the Angoff standard-setting process, the same panel reviewed the incorrect options and made judgements regarding their safeness. Discussion of the safeness of responses considered the safeness of the decisions if undertaken at the point of graduation to practise in New Zealand (exit Year 6, enter post-graduate Year 1). The responses were classified as not unsafe, and low, moderate, or high unsafeness. Again, the judgements were undertaken independently. This range of options including ‘not unsafe’ reflects the range that will occur in practice. Assigning a number from 1 (‘not unsafe’) to 4 (‘high unsafeness’), the median responses from the clinicians was taken as the safeness level for each option. The median was used because these are ordinal data, and this enabled results to be arranged simply on the scale from ‘not unsafe’ to ‘high unsafeness’. If the median was between categories, the more unsafe level was used.

6.2.2.4 Question selection for use in the research tests

From the 29 question pool, 24 questions were chosen for use. Selection was based on a balance of content and ensuring that all questions had at least one incorrect response deemed unsafe by the panel.
The 24 questions were then divided into two groups of 12, primarily balancing content, and again balancing estimated pass-mark and numbers of possible unsafe responses as far as possible.

Group A questions had an estimated pass-mark of 70.0%. Of the 96 incorrect responses, 24 were classified as not unsafe, 16 low unsafeness, 17 moderate unsafeness, and 39 high unsafeness. With 10 options per MCQ, the random chance of responding correct was 0.10, incorrect not unsafe 0.2, incorrect low unsafe 0.13, incorrect moderate unsafe 0.14, incorrect high unsafe 0.33, and 'don’t know' 0.10.

Group B questions had an estimated pass-mark of 64.9%. Of the 96 incorrect responses, 34 were classified as not unsafe, 25 low unsafeness, 19 moderate unsafeness, and 18 high unsafeness. With 10 options per MCQ, the random chance of responding correct was 0.10, incorrect not unsafe 0.28, incorrect low unsafe 0.21, incorrect moderate unsafe 0.16, incorrect high unsafe 0.15, and 'don’t know’ 0.10.

6.2.3 Paper development

Using the two sets of 12 questions (A and B), four papers were prepared. Each paper contained all questions, but the question group order, and the nature of the instructions and their order were varied. Order, instructions and content were balanced between papers (Table 6.1).
Table 6.1: Development of papers

<table>
<thead>
<tr>
<th></th>
<th>First 12-question test</th>
<th>Second-12 question test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First question group</td>
<td>Instruction</td>
</tr>
<tr>
<td>Paper 1</td>
<td>A</td>
<td>Confidence and safety</td>
</tr>
<tr>
<td>Paper 2</td>
<td>B</td>
<td>Number correct</td>
</tr>
<tr>
<td>Paper 3</td>
<td>A</td>
<td>Number correct</td>
</tr>
<tr>
<td>Paper 4</td>
<td>B</td>
<td>Confidence and safety</td>
</tr>
</tbody>
</table>

6.2.4 Confidence (certainty) descriptors

For each question, the students were asked to indicate their confidence in their response. As this test was completed on paper, this response could be made either after or before the question had been answered. The confidence level options that were provided were low, moderate and high. The descriptors provided to the candidates to guide them in determining their levels of confidence were categorised by perceived levels of knowledge and practice, including the readiness to ask for assistance as per Table 6.2.
<table>
<thead>
<tr>
<th>Related to knowledge</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Related to knowledge</td>
<td>The candidate has no idea of correct response, and any response would be a guess.</td>
<td>The candidate has no clear idea of correct response, but has some knowledge on the subject. Any response would be based on limited information.</td>
<td>The candidate has a reasonable idea of correct response on a basis of moderate knowledge on the subject. Any response would be based on sufficient information</td>
<td>The candidate is certain of correct response on a basis of detailed knowledge on the subject. Any response would not be a guess.</td>
</tr>
<tr>
<td>Related to practice</td>
<td>The candidate would need to consult a colleague or references prior to considering any response.</td>
<td>The candidate would need to consult a colleague or references, but would be able to give a response first.</td>
<td>The candidate would need to consult a colleague or references to confirm the correctness of the response.</td>
<td>The candidate would have no need to consult a colleague or reference.</td>
</tr>
</tbody>
</table>
6.2.5 Subjects: candidates, recruitment, consequence/stakes

Volunteers were sought by an email mail-out to all students in Years 2, 3, 4 and 5 in 2008. The email included information for students. All who agreed to participate gave written consent at the time of sitting the tests.

The tests did not contribute to any progress decisions, and students were free to withdraw at any stage.

6.2.6 Delivery format

The two 12-question MCQ tests were sat under approximate examination conditions (single invigilator in the room) in the evening, so as not to interfere with structured learning activities. The students sat the tests in teaching rooms, one at each of the three campuses, supervised by a single invigilator.

A test paper comprising two 12-question tests was given to each candidate. These were taken from a pile containing all four test papers in a random order. The papers included information and a consent form. There was no time limit for completion of the examination.

6.2.7 Information for candidates

The confidence and safety instruction and number correct instructions were different. The scoring system for the confidence and safety system was included (Table 6.3) with those instructions.

The following are the instructions given to the candidates:

Confidence and safety instruction

The scoring system in this part of the examination takes the candidate’s confidence and the consequences of the responses into account.

Each question has a correct answer. The incorrect answers have been classified as not unsafe, low level, moderate or highly unsafe. There is also a “don’t know” response.

We also want to take into account candidate confidence in responses. This will range from none, through low, moderate to high. There are
descriptors to reflect how these levels of confidence might relate to levels of knowledge and practice.

This scoring system is NOT the same as that currently used in the Year 5 MCQ examination.

For each question, mark one response square only.

*Number correct instruction*

The scoring system in this examination is +1 for a correct response and 0 for an incorrect response and also 0 for a “don’t know” response.

There is no penalty for incorrect responses.

This number correct scoring system is the same as that currently used in the Year 5 MCQ examination.

For each question, mark one response square only.

As has already been reviewed, confidence-based assessment incorporates the confidence of a candidate in the response, with several different scoring functions possible. For one of these, the positive and negative weightings of scores are constructed so that a negative score for an incorrect response is given greater weighting than the equivalent positive score for the same degree of confidence, such that the score is maximised if the candidate is accurately self-calibrated. These confidence-based scoring systems have been used recently, including with medical student assessments. However, none of these scoring functions incorporates different weightings for incorrect responses, such as for the degrees of unsafeness of the various responses. Hence the scoring grid was developed with the penalties for incorrect responses being incrementally increased as the level of unsafeness increased at each level of confidence (Table 6.3). The confidence and safety scoring system was presented to the participants in the form of a table that was included with the question instructions.
### Table 6.3: Confidence and safety scoring system

<table>
<thead>
<tr>
<th>Confidence</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>–</td>
<td>+1</td>
<td>+2</td>
<td>+3</td>
</tr>
<tr>
<td>Incorrect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not unsafe</td>
<td>–</td>
<td>0</td>
<td>-2</td>
<td>-6</td>
</tr>
<tr>
<td>Low unsafe</td>
<td>–</td>
<td>-1</td>
<td>-4</td>
<td>-10</td>
</tr>
<tr>
<td>Moderately unsafe</td>
<td>–</td>
<td>-2</td>
<td>-6</td>
<td>-14</td>
</tr>
<tr>
<td>Highly unsafe</td>
<td>–</td>
<td>-3</td>
<td>-8</td>
<td>-18</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

#### 6.2.8 Response format

The students responded with their chosen confidence level next to the correct answer selected. The second column contained the response options as shown in Figure 6.1.
<table>
<thead>
<tr>
<th>Degree of confidence in correct answer</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I do not know the answer and would look it up or ask the appropriate person for help</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6.1: Response options*

### 6.2.9 Feedback to student

Students received feedback on the numbers of their responses that were correct, incorrect, 'don’t know' and unsafe. They also received a number-correct score equivalent to the Year 5 exit standard. No normative data were included in the feedback.”
6.3 Analysis

6.3.1 Response and scoring analysis

Numbers of students and responses by paper number (instruction method, question pool, order) were recorded.

The responses to standard instructions were scored by number-correct scoring, i.e. +1 for a correct response and 0 for an incorrect response or a 'don’t know' response. The responses to confidence and safeness instructions were scored using the confidence and safeness scoring grid.

For this study, the response score reliability was calculated for a variety of parameters using Cronbach’s α. Following this, descriptions of mean scores and the range of scores were calculated for each of the scoring systems.

6.3.2 Effect of confidence and safeness instructions

It was envisaged that any effect of the order of presentation of the instructions on the outcomes (number of correct, incorrect, unsafe and ‘don’t know’ responses) would be reduced by the randomised approach with balanced papers. The effect of the order of the papers was analysed using independent sample t-tests in SPSS (SPSS Inc., Chicago, Illinois, USA).

Analyses were undertaken to address some of the research questions for this study.

Does including confidence (certainty)-in, and safeness-of, responses in assessment instructions alter students’ responses?

The effects of scoring instructions and student year group on the outcomes (number of correct, incorrect, unsafe and 'don’t know' responses) were analysed using repeated measures ANOVA in SPSS. This took account of both within subject (scoring instructions) and between subject (year group) effects. Interaction terms were tested to see if the effect of instruction depended on the student’s year group.

How can confidence (certainty)-in, and safeness-of, responses be incorporated into a scoring rubric?
Responses to the confidence and safeness instruction questions were scored using the rubric from Table 6.3 and also number-correct scoring (based on +1 for correct, 0 for incorrect and 0 for don’t know) irrespective of certainty and safeness. Mean scores and score reliability (Cronbach’s α) were calculated using SPSS for the number-correct scoring and confidence and safeness scoring system (Table 6.3).

Safeness score reliability (Cronbach’s α) was calculated based on 0 for don’t know, correct, and not unsafe, +1 for low level unsafeness, +2 for moderate level unsafeness, and +3 for high level unsafeness.

As an increasing score range can increase Cronbach’s α, analysis for this parameter was done by year group.

*Can we measure students’ correctness of responses for levels of confidence (certainty)?*

For each degree of confidence (low, moderate, high), the proportions of all students’ correct responses or incorrect, but not unsafe, responses were calculated and compared using ANOVA in SPSS, where the outcome variable was proportion of responses and covariates were level of confidence (certainty).

*Can we measure students’ safeness of responses for levels of confidence (certainty)?*

The proportion of all responses and of incorrect responses that were unsafe were calculated for each degree of confidence (low, moderate, high), and compared using ANOVA in SPSS, where the outcome variable was proportion of responses and covariates were level of confidence (certainty). In addition, the total numbers of responses that were unsafe by individual students were recorded.

*How do these measures of correctness and safeness for confidence (certainty) relate to students’ levels of experience?*

For each level of confidence (low, moderate and high), the mean proportion of students’ correct responses or incorrect but not unsafe responses was calculated and compared using ANOVA in SPSS, where the outcome variable was proportion of responses and covariates were level of confidence (certainty) and year group.
How do these measures of correctness and safeness for confidence (certainty) relate to students’ levels of knowledge?

Number correct was used to divide the students into four groups of ability, with number correct 0-3, 4-6, 7-9, and 10-12. As a measure of the effect of knowledge on confidence and safeness of responses, the proportions of any particular response (incorrect that were unsafe with levels of unsafeness; confidence) were compared by total number of correct responses, by ANOVA in SPSS, where the outcome variable was proportion of responses and covariates were level of confidence (certainty) and ability group.
6.4 Results

First to be described are the participants and their totalled responses. Then the analysis of the outcome regarding internal consistency of responses and scores. The next sections relate to results related to the effect of instructions on the responses, and finally analyses of various responses for levels of confidence (certainty).

6.4.1 Participants

In total, 374 students participated. In total 372/975 (38.1%) students from Years 2-5 participated. A number of students did not complete the certainty responses (one for student for Paper 1; two for Paper 2; one for Paper 3; and one for Paper 4). These results were also excluded from the analysis. Analyses were completed on 367 students, 94/251 (37.5%) of Year 2, 90/257 (35.0%) of Year 3, 51/229 (22.2%) of Year 4 and 132/238 (55.5%) of Year 5.

Responses

Of the 4404 question responses that related to the certainty and safety instructions, 1901 (43.1%) were correct, 1969 (44.7%) were incorrect, and 534 (12.1%) were 'don’t know'; 1016 (23.0%) were answered with high confidence, 1342 (30.5%) with moderate confidence, and 1512 (34.3%) with low confidence (Table 6.4).
Table 6.4: Total number (%) of responses by confidence and safeness under confidence and safeness instructions

<table>
<thead>
<tr>
<th>Response</th>
<th>Confidence</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td></td>
<td>–</td>
<td>480</td>
<td>642</td>
<td>779</td>
<td>1901</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10.9%</td>
<td>14.6%</td>
<td>17.7%</td>
<td>43.2%</td>
</tr>
<tr>
<td>Incorrect, not unsafe</td>
<td></td>
<td>–</td>
<td>498</td>
<td>338</td>
<td>79</td>
<td>915</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11.3%</td>
<td>7.7%</td>
<td>1.8%</td>
<td>20.8%</td>
</tr>
<tr>
<td>Incorrect and unsafe Low</td>
<td></td>
<td>–</td>
<td>194</td>
<td>159</td>
<td>100</td>
<td>453</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.4%</td>
<td>3.6%</td>
<td>2.3%</td>
<td>10.3%</td>
</tr>
<tr>
<td>Incorrect and unsafe Mod</td>
<td></td>
<td>–</td>
<td>147</td>
<td>69</td>
<td>21</td>
<td>237</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.3%</td>
<td>1.6%</td>
<td>0.5%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Incorrect and unsafe High</td>
<td></td>
<td>–</td>
<td>193</td>
<td>134</td>
<td>37</td>
<td>364</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.4%</td>
<td>3.0%</td>
<td>0.8%</td>
<td>8.3%</td>
</tr>
<tr>
<td>Don’t know</td>
<td></td>
<td>534</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>534</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.1%</td>
<td></td>
<td></td>
<td></td>
<td>12.1%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>534</td>
<td>1512</td>
<td>1342</td>
<td>1016</td>
<td>4404</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.1%</td>
<td>34.3%</td>
<td>30.5%</td>
<td>23.1%</td>
<td>100%</td>
</tr>
</tbody>
</table>
6.4.2 Internal consistency of responses and scores

For the two sets of 12 certainty-in and safeness-of response questions, internal consistency was greatest for certainty and least for safety (Table 6.5).

Table 6.5: Internal consistency (Cronbach’s α) for confidence in response, safeness of responses, correctness of response, composite score and number

<table>
<thead>
<tr>
<th></th>
<th>Confidence response (confidence safety instructions)</th>
<th>Safety response (confidence safety instructions)</th>
<th>Correctness of response (confidence safety instructions)</th>
<th>Confidence safety system scoring (confidence safety instructions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papers 1&amp;2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y2</td>
<td>0.86</td>
<td>0.13</td>
<td>0.31</td>
<td>0.38</td>
</tr>
<tr>
<td>Y3</td>
<td>0.82</td>
<td>0.29</td>
<td>0.35</td>
<td>0.59</td>
</tr>
<tr>
<td>Y4</td>
<td>0.83</td>
<td>&lt;0.01</td>
<td>0.09</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Y5</td>
<td>0.81</td>
<td>0.38</td>
<td>0.37</td>
<td>0.18</td>
</tr>
<tr>
<td>Papers 3&amp;4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y2</td>
<td>0.73</td>
<td>&lt;0.01</td>
<td>0.05</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Y3</td>
<td>0.61</td>
<td>0.02</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>Y4</td>
<td>0.73</td>
<td>0.17</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Y5</td>
<td>0.76</td>
<td>0.04</td>
<td>0.43</td>
<td>0.39</td>
</tr>
</tbody>
</table>

For each question answer, confidence (certainty) was designated as 0, 1, 2, 3; safety as 0, -1, -2, -3; correctness of response as 1, 0. The total score for each response that takes account of both confidence and safety was determined as per Table 6.3.

The internal consistency of the responses given to confidence (certainty) and safety instruction was lowest for safeness, higher for correctness and highest for confidence (certainty). The scoring rubric did not lead to improved internal consistency over number correct scoring.

For students sitting Papers 1 & 2, the maximum score was +31, the minimum score was -66 and the mean score -4.26. For students sitting Papers 3 & 4, the
maximum score was +7, the minimum score was -50, and the mean score -6.55. Given the composite nature of these scores, any particular score averaged across many questions has limited meaning, as opposed to the scores for individual questions or a pattern of scores across questions.

6.4.3 Effect of paper and instruction order

Paper order did have an effect on responses. Comparing the 12 questions the students answered first with the 12 questions answered second showed no difference in the mean number answered correctly (5.3 vs. 5.2, p=0.30) or number that were unsafe (2.9 vs. 3.1, p=0.06). However there was a significant difference in the number answered incorrectly, according to paper order, with more answered incorrectly in the second paper (5.2 vs. 6.1, p<0.001) and fewer answered 'don’t know’ responses in the second paper (1.4 vs. 0.7, p<0.001). This pattern was seen for both sets of instructions, suggesting that any differences resulting from the order of presentation of the groups of questions were the result of the order of presentation of questions, and not of the instructions.

6.4.4 Effect of experience (student year group) and instructions

The year group of the candidate led to differences in responses in ways that might be expected (Table 6.6). Increasing year group led to more correct, fewer incorrect, fewer 'don’t know’ and fewer unsafe (all Fs > 3.8, all p<0.001).

The type of instructions also led to differences in responses: Compared with number correct scoring instructions, safety and confidence scoring instructions led to fewer incorrect (F = 29.4, p<0.001), and more ‘don’t know’ responses (F = 87.7, p<0.001), with the same number of correct (F = 1.4, p=0.23), and unsafe (F = 3.3, p=0.07) responses.
Table 6.6: Number and percentage of responses by year group and instructions

<table>
<thead>
<tr>
<th>Year group</th>
<th>Scoring instructions</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Don’t know</th>
<th>Any unsafe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Safety and confidence</td>
<td>263</td>
<td>604</td>
<td>261</td>
<td>357</td>
<td>1128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23.3%</td>
<td>53.5%</td>
<td>23.1%</td>
<td>31.7%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Number correct</td>
<td>317</td>
<td>716</td>
<td>95</td>
<td>355</td>
<td>1128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.1%</td>
<td>63.5%</td>
<td>8.4%</td>
<td>31.5%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Safety and confidence</td>
<td>355</td>
<td>561</td>
<td>164</td>
<td>298</td>
<td>1080</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.9%</td>
<td>51.9%</td>
<td>15.2%</td>
<td>27.6%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Number correct</td>
<td>383</td>
<td>626</td>
<td>71</td>
<td>330</td>
<td>1080</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35.5%</td>
<td>58.0%</td>
<td>6.6%</td>
<td>30.6%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Safety and confidence</td>
<td>296</td>
<td>259</td>
<td>57</td>
<td>137</td>
<td>612</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48.4%</td>
<td>42.3%</td>
<td>9.3%</td>
<td>22.4%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Number correct</td>
<td>284</td>
<td>299</td>
<td>29</td>
<td>147</td>
<td>612</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46.4%</td>
<td>48.9%</td>
<td>4.7%</td>
<td>24.0%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Safety and confidence</td>
<td>987</td>
<td>547</td>
<td>50</td>
<td>262</td>
<td>1584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62.3%</td>
<td>34.5%</td>
<td>3.2%</td>
<td>16.5%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Number correct</td>
<td>988</td>
<td>575</td>
<td>21</td>
<td>287</td>
<td>1584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62.4%</td>
<td>36.3%</td>
<td>1.3%</td>
<td>18.1%</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Safety and confidence</td>
<td>1901</td>
<td>1971</td>
<td>532</td>
<td>1054</td>
<td>4404</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.2%</td>
<td>44.8%</td>
<td>12.1%</td>
<td>23.9%</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Number correct</td>
<td>1972</td>
<td>2216</td>
<td>216</td>
<td>1119</td>
<td>4404</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44.8%</td>
<td>50.3%</td>
<td>4.9%</td>
<td>25.4%</td>
<td></td>
</tr>
</tbody>
</table>

For the outcome variables (correct, incorrect, don’t know, any unsafe), an interaction term was included in the model to test whether the effect of instruction type differed according to student year groups: these interactions were not significant for correct (F=1.6, p=0.2) and unsafe responses (F=0.7, p=0.6).
For incorrect (F=3.1, p=0.02) and ‘don’t know’ (F=15.6, p<0.001) these interactions were significant, suggesting the effects of instruction did differ across year groups, with increasing year groups associated with less of a reduction in ‘don’t know’ and an increase in incorrect responses.

6.4.5 Students’ correctness of responses by level of confidence

The students’ responses demonstrated that with increasing confidence, there were more correct and fewer incorrect responses (Table 6.4). The proportion of responses that were correct rose with confidence: 31.7% (480/1512) when chosen with low confidence, 47.8% (642/1342) when chosen with moderate confidence, and 76.7% (779/1016) when chosen with high confidence (ANOVA, p<0.001). In addition, the proportion of responses that were incorrect and not unsafe fell from 32.9% (498/1512) when chosen with low confidence, to 25.2% (338/1342) when chosen with moderate confidence, and to 7.8% (79/1016) when chosen with high confidence (ANOVA, p<0.001).

6.4.6 Students’ unsafeness of responses by level of confidence

The proportion of responses that were both incorrect and unsafe to any degree fell from 35.3% (534/1512) for low confidence responses, to 27.0% (362/1342) for moderate confidence, and to 15.6% (158/1016) for high confidence (ANOVA, p<0.001) (Table 6.4).

However, of those responses that were incorrect, the proportions that were unsafe were 51.7% (534/1032) for low confidence responses, 51.7% (362/700) for moderate confidence and 66.7% (158/237) for high confidence (ANOVA, p<0.001) (Table 6.4). Although high confidence responses were more likely to be correct and safe, these results show that when a high confidence response was incorrect, it was more likely that the incorrect response was also unsafe.

There were 122 candidates (32.8%) who had a response that was unsafe to any degree and held with high confidence. Twelve of the responses from one candidate were unsafe, with three of these being highly unsafe. In addition, there were 31 candidates who gave a total of 37 highly unsafe responses held with high confidence, and of those, five students this way more than once (Table 6.7).
Table 6.7: Number unsafe responses per student

<table>
<thead>
<tr>
<th>Number of responses</th>
<th>Number of students with high confidence in any unsafe responses</th>
<th>Number of students with high confidence in highly unsafe responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>138</td>
<td>336</td>
</tr>
<tr>
<td>1</td>
<td>127</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

There were a small number of students who had multiple highly unsafe responses held with a high degree of confidence.

6.4.7 Students’ correctness for level of confidence by students’ levels of experience

The proportion of responses correct increased with increasing experience from 23.3% (263/1128) for Year 2, to 31.0% (335/1079) for Year 3, to 48.3% (296/612) for Year 4 to 62.3% (987/1584) for Year 5 (ANOVA, p<0.001).

The proportion incorrect and not unsafe was not different among year groups, being 21.3% (241/1128) for Year 2, 24.3% (262/1079) for Year 3, 19.9% (122/612) for Year 4 and 17.9% (283/1584) for Year 5 (ANOVA, p=0.601).

Increasing correctness for confidence (certainty) was seen for all year groups (Table 6.8), and also with increasing correctness with increasing year group for any given level of confidence (certainty).

For all year groups, the proportion of all responses that were incorrect but not unsafe decreased with increasing level of certainty. There was no difference in proportion correct across year groups for low and moderate certainty responses. There was a difference across year groups for the high certainty responses, but there was not a consistent pattern of change with increasing year group.
Table 6.8: Effect of experience on correctness for confidence: % responses correct and % responses incorrect and not unsafe for different levels of confidence and year group

<table>
<thead>
<tr>
<th>Year</th>
<th>Responses correct for each level of confidence</th>
<th>Responses incorrect and not unsafe for each level of confidence</th>
<th>ANOVA, p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Year 2</td>
<td>127/537</td>
<td>86/234</td>
<td>50/96</td>
</tr>
<tr>
<td></td>
<td>23.7%</td>
<td>36.73</td>
<td>52.1%</td>
</tr>
<tr>
<td>Year 3</td>
<td>161/495</td>
<td>128/307</td>
<td>66/113</td>
</tr>
<tr>
<td></td>
<td>32.5%</td>
<td>41.7%</td>
<td>59.3%</td>
</tr>
<tr>
<td>Year 4</td>
<td>89/211</td>
<td>91/194</td>
<td>116/150</td>
</tr>
<tr>
<td></td>
<td>42.2%</td>
<td>46.9%</td>
<td>77.6%</td>
</tr>
<tr>
<td>Year 5</td>
<td>103/268</td>
<td>337/607</td>
<td>547/657</td>
</tr>
<tr>
<td></td>
<td>38.4%</td>
<td>55.5%</td>
<td>83.3%</td>
</tr>
<tr>
<td>ANOVA, p</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

ANOVA, p

<table>
<thead>
<tr>
<th>Year 2</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>181/537</td>
<td>56/234</td>
<td>10/96</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>33.7%</td>
<td>23.9%</td>
<td>10.4%</td>
<td></td>
</tr>
<tr>
<td>Year 3</td>
<td>160/495</td>
<td>85/307</td>
<td>17/113</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>32.3%</td>
<td>27.7%</td>
<td>15.0%</td>
<td></td>
</tr>
<tr>
<td>Year 4</td>
<td>64/211</td>
<td>44/194</td>
<td>14/150</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>30.3%</td>
<td>22.7%</td>
<td>9.3%</td>
<td></td>
</tr>
<tr>
<td>Year 5</td>
<td>92/268</td>
<td>153/607</td>
<td>38/657</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>34.3%</td>
<td>25.2%</td>
<td>5.8%</td>
<td></td>
</tr>
<tr>
<td>ANOVA, p</td>
<td>0.78</td>
<td>0.60</td>
<td>0.004</td>
<td></td>
</tr>
</tbody>
</table>

182
6.4.8 Students’ safeness of responses for level of confidence by students’ levels of experience

The proportion of all responses that were unsafe decreased with increasing experience: 31.6% (357/1128) of Year 2 incorrect responses were unsafe, decreasing to 27.6% (298/1079) for Year 3, 22.4% (137/612) for Year 4 and 16.5% (262/1584) for Year 5 (ANOVA, p<0.001).

For Year 2 and 3 students there was no difference in the proportion of incorrect responses that were unsafe responses across the different levels of confidence (certainty). However, for Year 4 and 5 students, as confidence (certainty) increased, there was a reduction in the proportions of incorrect unsafe responses (Table 6.9).

Table 6.9: Effect of experience on proportions of all responses that are unsafe for different levels of confidence and year group

<table>
<thead>
<tr>
<th>Year</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>ANOVA, p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 2</td>
<td>229/537</td>
<td>92/234</td>
<td>34/96</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>42.6%</td>
<td>39.3%</td>
<td>35.4%</td>
<td></td>
</tr>
<tr>
<td>Year 3</td>
<td>174/495</td>
<td>94/307</td>
<td>30/128</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>35.2%</td>
<td>30.6%</td>
<td>26.6%</td>
<td></td>
</tr>
<tr>
<td>Year 4</td>
<td>58/211</td>
<td>59/194</td>
<td>20/150</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>27.4%</td>
<td>30.4%</td>
<td>13.3%</td>
<td></td>
</tr>
<tr>
<td>Year 5</td>
<td>73/268</td>
<td>117/607</td>
<td>72/657</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>27.2%</td>
<td>19.2%</td>
<td>11.0%</td>
<td></td>
</tr>
</tbody>
</table>

ANOVA, p

The proportion of incorrect responses that were unsafe varied with year group. The number incorrect and unsafe decreased with increasing experience. Students with more experience, those in Year 4 and Year 5, showed increasing confidence was associated with a reduced proportion of unsafe responses.
6.4.9 Students’ correctness of responses for level of confidence by students’ levels of knowledge?

All groupings by number correct (a surrogate for levels of knowledge) led to an increase in the proportion correct with increasing degrees of confidence (Table 6.9). Even the lowest scorers demonstrated a significant increase in the proportion correct as confidence (certainty) increased.

All groupings by number correct led to a decrease in the proportion incorrect but not unsafe with increasing degrees of confidence (Table 6.10). Even the lowest scorers demonstrated a significant decrease in the proportion incorrect but not unsafe as confidence (certainty) increased.

There was no difference in proportion correct or incorrect but not unsafe between groupings by number correct for each level of confidence (certainty). The differences in proportion incorrect but not unsafe for moderate confidence responses was at the border of statistical significance.
Table 6.10: Effect of knowledge on % responses correct and % responses incorrect and not unsafe for different levels of confidence and total correct

<table>
<thead>
<tr>
<th>For students with total correct of:</th>
<th>% correct and held with:</th>
<th></th>
<th></th>
<th>ANOVA, p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low confidence</td>
<td>Moderate confidence</td>
<td>High confidence</td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>148/497</td>
<td>225/428</td>
<td>264/330</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>29.8%</td>
<td>52.6%</td>
<td>80.0%</td>
<td></td>
</tr>
<tr>
<td>4-6</td>
<td>157/487</td>
<td>207/445</td>
<td>255/341</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>32.2%</td>
<td>46.5%</td>
<td>74.8%</td>
<td></td>
</tr>
<tr>
<td>7-9</td>
<td>152/452</td>
<td>179/407</td>
<td>234/313</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>33.6%</td>
<td>44.0%</td>
<td>74.8%</td>
<td></td>
</tr>
<tr>
<td>10-12</td>
<td>23/75</td>
<td>31/62</td>
<td>26/32</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>30.1%</td>
<td>50.0%</td>
<td>81.3%</td>
<td></td>
</tr>
<tr>
<td>ANOVA, p</td>
<td>0.63</td>
<td>0.082</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>For students with total correct of:</th>
<th>% incorrect but not unsafe and held with:</th>
<th></th>
<th></th>
<th>ANOVA, p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low confidence</td>
<td>Moderate confidence</td>
<td>High confidence</td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>155/497</td>
<td>85/428</td>
<td>21/330</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>31.2%</td>
<td>19.9%</td>
<td>6.4%</td>
<td></td>
</tr>
<tr>
<td>4-6</td>
<td>163/487</td>
<td>120/445</td>
<td>23/341</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>33.5%</td>
<td>26.9%</td>
<td>6.7%</td>
<td></td>
</tr>
<tr>
<td>7-9</td>
<td>157/452</td>
<td>118/407</td>
<td>31/313</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>34.7%</td>
<td>29.0%</td>
<td>9.9%</td>
<td></td>
</tr>
<tr>
<td>10-12</td>
<td>22/75</td>
<td>15/62</td>
<td>4/32</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>29.3%</td>
<td>24.2%</td>
<td>12.5%</td>
<td></td>
</tr>
<tr>
<td>ANOVA, p</td>
<td>0.60</td>
<td>0.016</td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>
6.4.9.1 Students’ unsafeness of responses and levels of confidence by students’ levels of experience?

There was no difference in the proportion of incorrect responses that were unsafe for Years 2, 3 or 4. However, for the Year 5 students there was a significant increase in the proportion of incorrect responses that were unsafe as confidence increased (Table 6.11).

For those students with the greatest experience, when they were highly confident in a response and it was incorrect, it was additionally more likely to be unsafe.

There was also a significant difference in the proportion of moderate confidence incorrect responses that were unsafe, with a reduction with increasing year group.

Table 6.11: Effect of experience on proportions of incorrect responses that were unsafe for different levels of confidence and year group

<table>
<thead>
<tr>
<th>Year</th>
<th>Confidence</th>
<th>ANOVA, p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Year 2</td>
<td>229/410</td>
<td>92/148</td>
</tr>
<tr>
<td></td>
<td>55.9%</td>
<td>62.2%</td>
</tr>
<tr>
<td>Year 3</td>
<td>174/334</td>
<td>94/179</td>
</tr>
<tr>
<td></td>
<td>52.1%</td>
<td>52.5%</td>
</tr>
<tr>
<td>Year 4</td>
<td>58/122</td>
<td>59/103</td>
</tr>
<tr>
<td></td>
<td>47.5%</td>
<td>57.3%</td>
</tr>
<tr>
<td>Year 5</td>
<td>73/165</td>
<td>117/270</td>
</tr>
<tr>
<td></td>
<td>44.2%</td>
<td>43.3%</td>
</tr>
<tr>
<td>ANOVA, p</td>
<td>0.061</td>
<td>0.001</td>
</tr>
</tbody>
</table>

186
6.4.9.2  Students’ response unsafeness and confidence by students’ levels of knowledge?

For one of the groupings by score, those who scored 4-6/12 correct, showed a significant increase in the number of incorrect and unsafe responses with increasing confidence. For all other groups there was no significant difference between levels of confidence in the proportion of incorrect responses that were unsafe. When incorrect with high confidence, those with a lower, but not lowest, number correct were more likely to be unsafe. For each level of confidence, there were no differences between groupings by number correct in the proportion incorrect and unsafe (Table 6.12).

Table 6.12: Effect of knowledge on % incorrect responses that were unsafe for different levels of confidence and total correct

<table>
<thead>
<tr>
<th>For students with total correct of:</th>
<th>% of incorrect that are unsafe held with:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low confidence</td>
<td>Moderate confidence</td>
<td>High confidence</td>
<td>ANOVA, p</td>
</tr>
<tr>
<td>0-3</td>
<td>194/349  55.6%</td>
<td>118/203  58.1%</td>
<td>45/66  68.1%</td>
<td>0.16</td>
</tr>
<tr>
<td>4-6</td>
<td>167/330  50.6%</td>
<td>118/238  49.6%</td>
<td>63/86  73.3%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>7-9</td>
<td>143/300  47.7%</td>
<td>110/228  48.3%</td>
<td>48/79  60.8%</td>
<td>0.10</td>
</tr>
<tr>
<td>10-11</td>
<td>30/52   57.7%</td>
<td>16/31   51.2%</td>
<td>2/6   33.3%</td>
<td>0.51</td>
</tr>
<tr>
<td>ANOVA, p</td>
<td>0.17</td>
<td>0.18</td>
<td>0.11</td>
<td></td>
</tr>
</tbody>
</table>
6.5 Discussion

Initially the results will be summarised in the context of the hypotheses and research questions. The results will be discussed related to feasibility of certainty-in and safeness-of responses, correctness for levels of confidence (certainty), and safeness for levels of confidence (certainty). Following this findings not related to the research questions are discussed. Lastly, the strengths and limitations of this study are discussed.

As a summary of the results for the hypotheses:

- Including confidence (certainty)-in, and safeness-of, responses in an MCQ assessment being feasible was confirmed.
- An item-scoring system for confidence (certainty)-in, correctness of, and safeness-of, responses being feasible, and possible to aggregate in a meaningful way was refuted.
- The information provided by including correctness for levels of confidence (certainty) and safeness for levels of confidence (certainty) being of greater value than the information provided by correctness of responses alone was probably confirmed, more information required.

As a summary of the results for the research questions:

1. How can confidence (certainty) and safeness be included in question construction and delivery?
   - Confidence in responses can be recorded by students using a scale linked to self-monitoring activities.
   - Safeness of incorrect responses can be stratified using the collective judgements of a panel of experts.

2. Does including confidence (certainty)-in, and safeness-of, responses in assessment instructions alter students’ responses?
   - Safety and confidence scoring instructions led to fewer incorrect responses.
   - Safety and confidence scoring instructions led to a greater change in responses for students in their first years of the course.
3. How can confidence (certainty)-in, and safeness-of, responses be incorporated into a scoring rubric?
   - The scoring rubric developed with aggregated scores increasing, based on a logarithmic scale, was not meaningful.
   - Confidence/correctness and confidence/safeness need to be considered as contributing to different ‘scores’.

4. Can we quantify students’ correctness of responses for levels of confidence (certainty)?
   - An increase in correctness for increasing levels of confidence was demonstrated.

5. Can we quantify students’ safeness of responses for levels of confidence (certainty)?
   - A decrease in unsafe responses for increasing levels of confidence was demonstrated.

6. How do these measures of correctness and safeness for levels of confidence (certainty) relate to students’ levels of experience?
   - An increase in correctness for increasing levels of confidence was demonstrated by students in all year groups.
   - A decrease in unsafe responses for increasing levels of confidence was demonstrated by students in the highest year group.
   - Of the incorrect responses, a greater proportion of high confidence responses were unsafe for the highest year group.

7. How do these measures of correctness and safeness for levels of confidence (certainty) relate to students’ levels of knowledge?
   - An increase in correctness for increasing levels of confidence was demonstrated by students in all knowledge ability groups.
   - Of the incorrect responses, those with high confidence were more likely to be unsafe for students in the lower, but not lowest, knowledge group.
6.5.1 Feasibility of certainty-in and safeness-of responses

Combining correctness, safeness and levels of certainty into a single composite score was unsuccessful. Even though scoring rubrics, as originally described \(^{16,38}\), were used as a basis to generate scores for each candidate to reflect levels of correctness, certainty and safety, this did not work as envisaged. The meaning of an average of a composite score across several questions is limited. What is the meaning of a year group mean score or -4 or -6? What is the meaning of an individual’s score of +31, +7, -66 or -50?

The composite score reliability was also lower than for a number correct, being less than 0.01 for three of year groups across both papers, compared to the reliability of the correctness of responses of less than 0.01 for one year group. The multiple contributing factors to scores meant that scores lost their meaning. There are two reasons why this response reliability and score reliability pattern could have happened. Importantly, aggregation across the characteristics such as knowledge/ignorance, safety/danger, confidence/doubt would be less meaningful and not support the score interpretations, as validation, if these attributes are not part of an overall attribute \(^{127}\). Secondly, even if there is an overall attribute, the size of the penalty for a potential miss-key error, where a candidate fills in the question response or certainty level incorrectly, might also not support the score interpretation, as validation.

Estimating a tendency for error, as being unsafe, is possible by counting incorrect responses, provided there is the opportunity to respond with a DKO \(^{31}\). There are different options to aggregate correct and incorrect responses \(^{127}\), including the correct score having to be above a threshold, and the incorrect score having to be below a threshold.

What results representation has more meaning? A count of responses in various correct/incorrect/safeness/confidence (certainty) categories, with requirements for each needed to pass may be more appropriate. One such scheme is shown in Table 6.13. This recognises the importance of correct knowledge that is held with appropriate confidence (certainty). Also, it means that correct responses cannot be used to compensate for unsafe responses, especially those held with high confidence (certainty).
By considering the combination of safeness and confidence (certainty), a risk table could be produced (Table 6.13).

As an example, to pass an assessment, these four criteria might need to be met:

1. The total number of correct responses (total of cells a) has to be above a threshold.
2. The total number of unsafe responses held high confidence (certainty) unsafe responses (cells c+e) has to be below a threshold.
3. The total number of highly unsafe responses (cells d+e) has to be below a threshold.
4. The total number of highly unsafe responses held with high confidence (certainty) (cells e) has to be below a threshold.

All responses cells a-e would be useful to guide future learning.

Table 6.13: An example of combined confidence (certainty) and safety scoring used to make summative decisions

<table>
<thead>
<tr>
<th></th>
<th>Nil</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not unsafe</td>
<td>-</td>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>Low unsafe</td>
<td>-</td>
<td>b</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>Moderately unsafe</td>
<td>-</td>
<td>b</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>Highly unsafe</td>
<td>-</td>
<td>d</td>
<td>d</td>
<td>e</td>
</tr>
<tr>
<td>Don’t know</td>
<td>b</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
6.5.2 Correctness for levels of confidence (certainty)

In this study increasing confidence (certainty) was associated with more correct, and fewer incorrect responses: the students generally knew if they were likely to be correct. Even those with least experience and fewer numbers correct had increasing correctness for levels of confidence (certainty). This demonstration of a self-awareness even for those with lower scores and lower levels of experience is contrary to the generally accepted notion that ‘the unskilled are unaware’.  

The general literature and healthcare literature on self-assessment tends to focus on people estimating their rank in a group, estimating their own score or grade, or their likelihood of being correct. The wording and format of the confidence (certainty) questions in this study were more authentic to clinical practice, which may explain the differences. Response scales work best when they reflect cognitive structuring, which in this case was targeted to self-monitoring as reflection-in-action and the need to consult for assistance, specifically, the wording referred to when ‘to look it up’ or ‘defer to others’.  

This finding, similar to that of others, suggests that self-monitoring might be present for those with lower scores, albeit among selected participants.

6.5.3 Safeness for levels of confidence (certainty)

In this study, a substantial number of incorrect responses appeared to have some degree of unsafeness, and this occurred across all years. There was a small subset of students who showed a concerning pattern of providing unsafe responses held with high confidence (certainty). Increasing confidence (certainty) was associated with fewer incorrect and fewer unsafe responses. However, the proportion of incorrect responses that were unsafe did not fall, and in fact increased with increasing confidence (certainty).

The Year 5 cohort and the group of students with a number correct of 4-6/12 had a pattern of responding which could be concerning. When incorrect, as their certainty increased, their responses were more likely to be unsafe. If extrapolated to practice, such behaviour would be concerning, but it is probable that responses as an undergraduate in a research assessment might not be reflected in subsequent practice.
Why might the students make higher confidence (certainty) unsafe responses? Firstly, the students have only experienced number correct scoring with feedback that related to correctness or not, rather than to safety. Feedback that is verification without elaboration is less useful in guiding learning 154. Could it be that simply scoring on the number correct, which these students had experienced, encourages a ‘guess when unsure’ strategy, despite this being inappropriate in clinical practice 93? The students’ typical assessment behaviour could therefore tend towards giving responses that have a chance of being correct, because safety of incorrect responses has never before been raised as an issue.

Secondly, it is possible that ‘knowing what to do’ is learnt before ‘knowing what not to do’. Knowing what not to do can be termed negative knowledge 350, and in the current study, this would be awareness of responses to avoid because of unsafeness. On the path from novice to expert, negative knowledge tends to be gained by experience, rather than from standard learning resources, such as structured learning sessions. The development of negative knowledge, selecting choices that are less likely to lead to a negative outcome, may contribute to the increasing certainty of experts. However, those who are not yet expert may not have had sufficient experience to develop negative knowledge with appropriate confidence (certainty). These learners still choose unsafe responses with higher confidence (certainty).

Thirdly, it has been shown that as performance and confidence (certainty) increase, some people are more likely to take risks 224. Although this has not been shown for those in healthcare clinical practice or students specifically, if this was a contributory factor, this would be a significant finding with regard to training, and thus would require further exploration. Increasing ability may make some errors more likely, but the increase in ability is also associated with better error detection and correction 250. In other words, ability and self-monitoring may not increase in parallel: if a student’s improvement in self-monitoring lags behind their improvement in ability, they could go through a stage of increased ability and low self-monitoring before reaching a stage of increased ability and appropriate self-monitoring, akin to Pope’s (1688-1744) “A little learning is a dangerous thing; Drink deep, or taste not the Pierian spring”.

193
Lastly, some doctors, and therefore possibly those who are in training, will have a level of confidence (certainty) in their applied practice that is greater than that supported by their ability. This may be due to doctors not disclosing, and therefore ignoring, uncertainty, so perceived certainty is increased. There are many heuristics and biases in clinical practice that normally function to aid clinical practice, appropriately increasing correctness and therefore confidence (certainty), but these can at times lead to incorrect decisions being made with undue confidence (certainty).

6.5.4 Finding not related to research questions

The internal consistency of the confidence (certainty) responses was considerable higher than those for correctness and safeness. This would not be unexpected for confidence (certainty) and correctness as it has been reported in the prior literature. Given, the internal consistency of the composite scoring system was also limited, and as discussed above the composite scoring system was not successful, this finding related to internal consistency of response is of interest, but cannot be extrapolated further. These analyses will be reviewed in the discussion sections of subsequent studies and also related to all studies in this thesis in section 10.3.2.

6.5.5 Strengths of this study

A strength of this study is the randomised crossover methodology. This should reduce the bias caused by order of questions or order of scoring instructions. Many year groups were represented and as expected, the responses changed with increasing year group, supporting validity of the questions. Another strength is that the descriptors of confidence were linked to clinical practice in terms of contacting a colleague or referring to relevant literature, making this authentic.

The distractors were classified into degrees of unsafeness by a panel of practicing clinicians, thereby increasing the authenticity of the classifications. It is possible that the analysis of unsafeness of incorrect responses could be flawed by misjudgement in the allocation of the degree of unsafeness by the reviewing panel. However, the median judgement of several independent reviewers, all of whom were practicing clinicians, was used to try and mitigate this. Safety classifications based on actual data of patient outcomes may be possible.
but this is unrealistic for all distractors, and it would be impractical to limit
distractors to those where data are available. However, using evidential data that
were available, to complement and validate judgements by experts regarding
unsafeness, would improve this process.

The response pattern has similarities to those described recently by Dory et al.\textsuperscript{352},
with confidence-based marking being used to add to aid the identification of
unsafe gaps in knowledge. Unlike Dory et al.’s research\textsuperscript{352}, which used the percent
probability of being correct as a descriptor for confidence, and 'hazardous
ignorance' defined as incorrect response held with high confidence, the
categorisation of confidence (certainty) and safety used in the current study was
better related to actual practice. The descriptors of confidence (certainty) were
linked to practice in terms of contacting a colleague or referring to relevant
literature. Although an incorrect response held with confidence may be
hazardous, the content and context of the response is also important and needs to
be considered. The current study did take the context and context into account.

6.5.6 Limitations of this study

There are limitations to Study 1. All the participants were aware that this was not
a summative assessment and was part of a research project. Their responses may
have been different if this was a high-stakes examination or they were in clinical
practice. Question selection may not be representative, as there were many more
unsafe response options in this study than the average in the question bank.
Although the instructions did cause the participants to alter their responses, the
effect size was smaller than that between consecutive year groups.

A different standardisation point was chosen for the pass-mark (exit Year 5) and
safety judgement (exit Year 6). Different standardisation points were chosen in an
attempt to reduce rather than increase complexity. The panel members were
familiar with pass-mark setting for exiting Year 5, as this was the standard for the
Year 5 common component examination. The safety judgement of exiting Year 6
was chosen, as this is the time of starting postgraduate practice. Although Year 6
Trainee Interns do have a degree of practice as a member of a healthcare team, all
decisions they make are more closely supervised than for postgraduate Year 1
doctors\textsuperscript{353}.\textsuperscript{353}.
Are the unsafe responses truly unsafe? Basing the safety classification on clinical outcome data from authentic practice would be ideal, but such data are unlikely to be available for all distractors \(^{29}\). However, it is reassuring that safeness was classified by the judgements of a review panel, a method consistent with previous research defining dangerous responses from the majority decision of a reviewing panel \(^{240,241}\). Further analysis such as kappa analysis of the panellists’ judgements would add to the research.
Chapter 7. Study 2
7.1 Introduction

7.1.1 Building on prior studies

Building on Study 1, and prior research as outlined in the literature review, there are aspects that need to be developed and explored as part of Study 2.

Combining correctness, safety and levels of certainty to produce a single score did not work from a psychometric or a theoretical perspective. However, correctness, safeness and certainty are important issues that need to be considered. For Study 2, a means of combining the information from correctness and safeness of responses, and certainty in responses needed to be investigated.

The research test in Study 1 had no stakes attached. For Study 2, undertaking certainty-in and safeness-of MCQ responses in an assessment with higher stakes also needed to be investigated. A surprising result from Study 1 was that the proportion of incorrect responses that were unsafe did not fall with increasing certainty. In study 2, this will be explored further.

7.1.2 Hypotheses and research questions

The hypotheses for Study 2 became:

- Undertaking certainty-in and safeness-of MCQ responses as a format for an assessment used to inform high-stakes decisions would be feasible.
- Students of all ability levels would demonstrate increasing correctness with certainty.
- Unsafe incorrect responses would be held with increasing certainty by those with increasing levels of ability.

The research questions for Study 2 became:

1. Can certainty-in and safeness-of response be included in an MCQ examination and used to inform high-stakes decisions?
2. What is the relationship between response certainty and odds of correctness?
3. What is the relationship between response certainty and the odds of potentially unsafe consequence responses?
4. How do these relationships of certainty with correctness and unsafeness vary with ability groupings, derived from the number correct on the assessment?
7.2 Methods

7.2.1 Design

The design chosen was to analyse and report the results of assessment responses by Year 5 students for an MCQ assessment incorporating certainty and safety, used to inform progression decisions.

The Medicine degree course at the University of Otago is described in section 4.3. The module in which the assessments informed progression decisions was Year 5 General Medicine and Subspecialties at University of Otago, Wellington, between 2011 and 2015.

7.2.2 Question pool content and review

To create a question bank for the MCQ examination, 43 new questions were created. Questions were selected to reflect the topics of General Medicine and the respiratory, cardiology, oncology, and nephrology subspecialties, as well as outcomes for the course.

Some previously-used questions were modified with the addition of a variable number of distractors, such that all responses had 10 possible responses with the same last option for all questions being “I do not know, so would ask someone or seek a reference”. Some new questions were created, again with 10 response options including a last DKO.

A panel of six hospital-based senior physicians, who were also university staff members, reviewed question content and suggested changes, and questions were developed. Following this the questions were finalised, and the correct answer confirmed.

7.2.3 Standard-setting process

The standard-setting process used was the Modified Angoff. This process is used for standard-setting for the OMS Year 5 common component exam, and the same six member panel was used as the panellists, being familiar with the process.

The standard to be used to produce judgements was “minimum to exit Year 5 and enter Year 6 of the course”. Discussion ensured that the panellists had a common
understanding. Independently, each panellist gave a judgement of the percentage of students with a performance standard of “minimum to exit Year 5 and enter Year 6 of the course” who would get the question correct. This was done using a Likert scale of 0-100%, with 10% increments. The mean judgement was used to create the estimated pass-mark for each question.

7.2.4 Question classification for safety process

Simultaneously with the Angoff standard-setting process, the same six physicians reviewed the incorrect options and made judgements regarding safeness. Discussion of the safeness of responses for each option considered the safety of the decisions if undertaken at the point exiting Year 5 and entering the Year 6 Trainee Intern year. The responses were classified as not unsafe, and low, moderate, or high unsafeness. Again, the judgements were undertaken independently. This range of options including 'not unsafe' reflects the range that will occur in practice. Assigning a number from 1 ('not unsafe') to 4 ('high unsafeness'), the median responses from the clinicians was taken as the safeness level for each option. The median was used because these are ordinal data. If the median was between categories, the more unsafe category was used.

Of the 344 incorrect responses, 135 were classified as not unsafe, 127 low unsafeness, 68 moderate unsafeness, and 14 high unsafeness. With 10 options per MCQ, the random chance of responding correctly was 0.10, incorrect not unsafe 0.31, incorrect low unsafe 0.30, incorrect moderate unsafe 0.16, incorrect high unsafe 0.03, and 'don't know' 0.10.

The 43 questions each had eight incorrect responses, giving a total of 344. Each panellist did not review every incorrect response, but the mean number of reviews per incorrect response was 4.2 (range 2-6). Kappa statistic for agreement in these judgements was 0.29 using Stat IC 14.1 (StataCorp LP, College Station, Texas, USA).

7.2.5 Question selection for use in the examinations

From the 43 question pool, six examinations each of 20 questions were chosen. This was based on a balance of content and such that the pass-mark for each was 12/20 correct.
7.2.6 Subjects: candidates, recruitment, consequence/stakes

The students, whose assessment responses were the focus for this study, had all progressed from Year 4 to Year 5 at one campus. Each year, it is assumed likely that the level of performance within the group will range from ‘just below the standard’ required to exit Year 5, through to ‘excellent’ for the year of study. Decisions to halt progress are usually not made at the end of the year, but instead are pre-empted by interim decisions made during the year in order to allow students time to make up for any deficiencies in their assessed performance during the course of the year.

7.2.7 Delivery format

The assessment process was explained to the students at the start of the module by a member of the academic staff. In addition, written information and practice questions were placed on the students’ learning management system (LMS).

Questions were delivered using the LMS. The front page of the assessment included a reminder to the students of the descriptors for levels of certainty (Table 7.1). These descriptors were developed from those used in Study 1. To respond to the questions, the students clicked one check box for their response choice and, unless the DKO (corresponding to no certainty) response was selected, a second check box was provided in which they could indicate their level of certainty (low, moderate or high).

The time allowed was 40 minutes to answer 20 questions, in a computer lab with an invigilator present.

The examinations usually occurred on the last Wednesday afternoon of each five-week module.
<table>
<thead>
<tr>
<th>Certainty</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related to knowledge</td>
<td>The student has no idea of the correct response and any response would be a guess.</td>
<td>The student has no clear idea of the correct response but has some knowledge on the subject. Any response would be based on limited information.</td>
<td>The student has a reasonable idea of the correct response on a basis of moderate knowledge on the subject. Any response would be based on sufficient information</td>
<td>The student is certain of the correct response on a basis of detailed knowledge on the subject. Any response would not be a guess.</td>
</tr>
<tr>
<td>Related to practice</td>
<td>The student would need to consult a colleague or references prior to considering any response.</td>
<td>The student would need to consult a colleague or references, but would be able to give a response first.</td>
<td>The student would need to consult a colleague or references to confirm the correctness of the response.</td>
<td>The student would have no need to consult a colleague or reference.</td>
</tr>
</tbody>
</table>

*Table 7.1: Certainty descriptors*
7.2.8 Information for students

In addition to providing feedback to the students for their use in future learning, the response pattern was used for a summative purpose to inform pass/fail decisions for the module. Information of certainty-in and safeness-of responses was used to make these decisions. Therefore, the pass/fail decision was not based solely on number correct. The students were aware of the pass/fail criteria in advance.

The following constituted a fail:

- More than 2/20 highly or moderately unsafe responses that were held with high certainty, irrespective of the number correct.
- A score of fewer than 8/20 correct.

The following constituted a pass:

- A score of 12/20 or more correct, and not meeting any fail criteria.

The following constituted an excellent pass:

- A score more than 15/20 correct, with three or more correct responses held with high certainty, and not meeting any fail criteria.

A score of 8-11/20 could be tolerated, allowing the student time to improve, unless information from other assessments within the module indicated a global substandard performance.

These levels of classification and tolerance were decided by a consensus discussion by module staff.

7.2.9 Feedback to students

Feedback sent to students was in the form of a matrix, displaying for each question, the question topic, the certainty of their response and safeness. In addition, specific notice was drawn to the topics of those questions with incorrect responses higher in certainty and unsafeness. No normative data were included in the feedback.
7.3 Analysis

7.3.1 Response and scoring analysis

With regard to general data analysis, reliability as internal consistency for correct responses was calculated, irrespective of certainty. From this, and the standard deviation of total correct for each student, the standard error of measurement (SEM) for number correct was calculated. Reliability, as internal consistency of certainty and safeness of responses, was also calculated.

7.3.2 Analysis linked to research questions

To address each of the research questions, the following statistical analyses were undertaken.

For each research question, a mixed-model logistic regression analysis was used which controlled for module group and student by including those as random terms. To examine the relationship between being correct and response certainty the log of the odds of being correct (vs. not) was modelled with terms for level of certainty (as categories). For the relationship between a potentially unsafe consequence and response certainty, both the log of the odds of any unsafe response (vs. not) and the log of the odds of a moderately or highly unsafe response (vs. low or not unsafe) were modelled with terms for level of certainty (as categories). The relationships with unsafe responses were examined for all (correct and incorrect) responses, and among incorrect responses. The models included random terms for level of certainty and student module group with students as subjects nested in the module groups. To test whether these relationships with certainty and consequence vary with score groupings, additional models for each outcome were analysed with terms for these score groups (as categories), and their interactions with level of certainty were added to the models. When interactions were significant, further testing was undertaken regarding whether there were differences between levels of certainty within each group, and whether differences between levels of certainty were different between groups. A p-value <0.05 was considered significant.

When there were significant differences between the levels of certainty, comparisons between levels of certainty (three comparisons) were adjusted for multiple comparisons with the Holm–Bonferroni method. When there was a significant interaction, tests for any differences between levels of certainty (four comparisons,
one for each score group), comparisons between levels of certainty within each significant score group (three comparisons), and comparisons of differences between level of certainty between score groups (six comparisons, between pairs of score groups) were also adjusted with the Holm–Bonferroni method. When these later tests were significant, differences between levels of certainty (three comparisons) were also adjusted. Each adjustment was made independently. The adjusted criteria for considering p-values significant were 0.05 divided by the number of comparisons for the smallest p-value, 0.05 divided by one less than the number of comparisons for the next smallest, and similarly for others with 0.05 divided by the number of comparison less the number already considered. Once a p-value was considered not significant, no further p-values were examined.

The GLIMMIX procedure of SAS 9.4 (SAS Institute Inc., Cary, North Carolina, USA) was used, with Laplace’s method and the between-subject and within-subject method for the denominator degrees of freedom.

The scores of 8/20, 12/20 and 16/20, which had been used for decision-making thresholds, having been generated from the Angoff procedure and consensus discussion amongst module staff as described above, were used to group the students by assessment performance. Score groupings were:

1. low scorers (<8/20 correct).
2. below-standard scorers (8-11/20 correct).
3. above-standard scorers (12-15/20 correct).
4. excellent scorers (>15/20 correct).
7.4 Results

Firstly, the participants and their totalled responses will be described, followed by the analysis of the outcome regarding internal consistency of responses. The next sections relate to analyses of various responses for levels of certainty and ability level.

7.4.1 General response data

Four students’ data had to be removed from the analysis because of technical problems with question delivery on the LMS. The remaining 330 students sat the examinations, with 20 students scoring in the low score group (SG1), 159 in the below-standard group (SG2), 133 in the above-standard group (SG3), and 18 in the excellent group (SG4).

Of the 6600 question attempts, 284 (4.3%) were answered 'don’t know' and 3,747 (56.8%) were answered correctly. Of the correct responses, 807 (21.5%) were low certainty, 1,675 (44.7%) moderate certainty and 1,265 (33.8%) high certainty.

Of the 2,569 incorrect responses 1,407 (54.8%) were not unsafe, 986 (38.4%) low level unsafe, 146 (5.7%) moderately unsafe and 30 (1.2%) highly unsafe; 930 (36.2%) were low certainty, 1,192 (46.4%) moderate certainty and 447 (17.4%) high certainty.

There were four responses held with high certainty deemed to be highly unsafe, each from a different student, one of whom was in the SG3, two from SG2 and one from SG1. No student had three or more highly or moderately unsafe responses that were held with high certainty.

The responses for each Score Group as shown (Table 7.2 – Table 7.5).
Table 7.2: Responses for students in Score Group 1

<table>
<thead>
<tr>
<th>Responses SG1</th>
<th>Certainty</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td></td>
<td>32</td>
<td>54</td>
<td>37</td>
<td>123</td>
</tr>
<tr>
<td>Incorrect, not unsafe</td>
<td></td>
<td>44</td>
<td>49</td>
<td>22</td>
<td>115</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>30</td>
<td>26</td>
<td>22</td>
<td>78</td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Don’t know</td>
<td>64</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
<td>116</td>
<td>133</td>
<td>87</td>
<td>400</td>
</tr>
</tbody>
</table>

Table 7.3: Responses for students in Score Group 2

<table>
<thead>
<tr>
<th>Responses SG2</th>
<th>Certainty</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td></td>
<td>304</td>
<td>661</td>
<td>592</td>
<td>1557</td>
</tr>
<tr>
<td>Incorrect, not unsafe</td>
<td></td>
<td>278</td>
<td>416</td>
<td>109</td>
<td>803</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>181</td>
<td>242</td>
<td>124</td>
<td>547</td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td>36</td>
<td>42</td>
<td>10</td>
<td>88</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Don’t know</td>
<td>167</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>167</td>
</tr>
<tr>
<td>Total</td>
<td>167</td>
<td>807</td>
<td>1369</td>
<td>837</td>
<td>3180</td>
</tr>
</tbody>
</table>
### Table 7.4: Responses for students in Score Group 3

<table>
<thead>
<tr>
<th>Responses SG3</th>
<th>Certainty</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td></td>
<td>–</td>
<td>410</td>
<td>801</td>
<td>559</td>
<td>1770</td>
</tr>
<tr>
<td>Incorrect, not unsafe</td>
<td></td>
<td>–</td>
<td>195</td>
<td>223</td>
<td>38</td>
<td>456</td>
</tr>
<tr>
<td>Incorrect and unsafe: Low</td>
<td></td>
<td>–</td>
<td>98</td>
<td>145</td>
<td>91</td>
<td>334</td>
</tr>
<tr>
<td>Incorrect and unsafe: Moderate</td>
<td></td>
<td>–</td>
<td>21</td>
<td>12</td>
<td>8</td>
<td>51</td>
</tr>
<tr>
<td>Incorrect and unsafe: High</td>
<td></td>
<td>–</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Don’t know</td>
<td></td>
<td>53</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>53</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>53</td>
<td>725</td>
<td>1185</td>
<td>707</td>
<td>2660</td>
</tr>
</tbody>
</table>

### Table 7.5: Responses for students in Score Group 4

<table>
<thead>
<tr>
<th>Response SG4</th>
<th>Certainty</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td></td>
<td>–</td>
<td>61</td>
<td>159</td>
<td>77</td>
<td>297</td>
</tr>
<tr>
<td>Incorrect, not unsafe</td>
<td></td>
<td>–</td>
<td>20</td>
<td>10</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>Incorrect and unsafe: Low</td>
<td></td>
<td>–</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>27</td>
</tr>
<tr>
<td>Incorrect and unsafe: Moderate</td>
<td></td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Incorrect and unsafe: High</td>
<td></td>
<td>–</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Don’t know</td>
<td></td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0</td>
<td>89</td>
<td>180</td>
<td>91</td>
<td>360</td>
</tr>
</tbody>
</table>

Each of the six tests consisted of 20 MCQ questions, with limited reliability as internal consistency (Cronbach’s $\alpha$) for correctness, and SEM for total correct between 8-10% (1.73-1.93/20), as shown in Table 7.6. The internal consistency
(reliability) of the responses for certainty was higher than for correctness, which was in turn higher than safeness.

Table 7.6: Measures of internal consistency of correct responses and variance in total correct across the 6 tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Number of students</th>
<th>Cronbach's α for correct responses</th>
<th>SD for total correct</th>
<th>SEM for total correct</th>
<th>Cronbach's α for certainty responses</th>
<th>Cronbach's α for safe responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>0.21</td>
<td>2.17</td>
<td>1.93</td>
<td>0.83</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>0.42</td>
<td>2.47</td>
<td>1.89</td>
<td>0.79</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>0.43</td>
<td>2.29</td>
<td>1.73</td>
<td>0.88</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>0.43</td>
<td>2.33</td>
<td>1.76</td>
<td>0.71</td>
<td>0.30</td>
</tr>
<tr>
<td>5</td>
<td>52</td>
<td>0.08</td>
<td>1.90</td>
<td>1.82</td>
<td>0.86</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>0.68</td>
<td>3.27</td>
<td>1.84</td>
<td>0.86</td>
<td>0.48</td>
</tr>
</tbody>
</table>

7.4.2 Relationship between levels of certainty and the odds of correctness

Overall students giving levels of certainty were more likely to be correct. The odds ratio (OR) for being correct was 1.7 (95%CI, 1.5–2.0) for moderate vs. low certainty, 2.1 (95%CI, 1.9–2.5) high vs. moderate certainty and 3.7 (95%CI, 3.1–4.3) for high vs. low certainty (all p<0.0001).

7.4.3 Relationship between levels of certainty and the odds of correctness by ability group

The levels of certainty were significantly different between score groups (interaction of score group and level of certainty F=3.6, degrees of freedom (df)=6570, p=0.002, Table 7.7). The odds of being correct among levels of certainty for all score groups were significantly different (p=0.045 for SG1 and p<0.0001 for each of the other groups). After adjusting for multiple comparisons, there were no significant differences between levels of certainty for SG1. There were significant differences between all levels of certainty for the other score groups, except for SG4 between moderate and high certainty (p=0.39).
After adjusting for multiple comparisons between groups, the ORs for SG2 were significantly different to those for SG4 (p=0.003). Other differences between score groups were not significant: SG1 compared to SG2, SG3 and SG4 (p=0.010, p=0.14 and p=0.29 respectively), SG2 compared to SG3 (p=0.14) and SG3 compared to SG4 (p=0.018).

Table 7.7: Odds ratio of a correct response by certainty and score group

<table>
<thead>
<tr>
<th>Score group</th>
<th>SG1 n=20</th>
<th>SG2 n=159</th>
<th>SG3 n=133</th>
<th>SG4 n=18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low certainty</td>
<td>1.0</td>
<td>1.6 (1.0–2.4)</td>
<td>3.4 (2.2–5.3)</td>
<td>5.7 (3.1–10.5)</td>
</tr>
<tr>
<td></td>
<td>p=0.036</td>
<td>p&lt;0.0001</td>
<td>p&lt;0.0001</td>
<td>p&lt;0.0001</td>
</tr>
<tr>
<td>Moderate certainty</td>
<td>1.8 (1.1–3.1)</td>
<td>2.5 (1.6–3.7)</td>
<td>5.5 (3.6–8.4)</td>
<td>19.9 (10.8–36.6)</td>
</tr>
<tr>
<td></td>
<td>p=0.032</td>
<td>p&lt;0.0001</td>
<td>p&lt;0.0001</td>
<td>p&lt;0.0001</td>
</tr>
<tr>
<td>High certainty</td>
<td>1.9 (1.1–3.5)</td>
<td>6.3 (4.1–9.8)</td>
<td>10.6 (6.8–16.7)</td>
<td>14.4 (7.2–29.1)</td>
</tr>
<tr>
<td></td>
<td>p=0.027</td>
<td>p&lt;0.0001</td>
<td>p&lt;0.0001</td>
<td>p&lt;0.0001</td>
</tr>
</tbody>
</table>

Odds ratio (95% confidence interval)

7.4.4 Relationship between levels of certainty and the odds of unsafeness among all (correct and incorrect) responses

Students were less likely to give an unsafe response with moderate or high certainty than low certainty. The OR for moderate certainty vs. low was 0.71 (95% CI, 0.61–0.83; p<0.0001), and for high vs. low, 0.65 (95% CI, 0.55–0.78; p<0.0001).

7.4.5 Relationship between levels of certainty and the odds of unsafeness among all (correct and incorrect) responses by ability group

The levels of certainty were significantly different between score groups (interaction of score group and level of certainty F=2.7, df=6570, p=0.013, Table 7.8). There were significant differences in the odds of an unsafe response between levels of certainty
for SG2 (p<0.0001, where all differences were significant), but not the other groups of scorers (SG1 p=0.095, SG3 p=0.20, SG4 p=0.25).

After adjusting for multiple comparisons, the differences in OR between groups were not significant, SG1 compared to SG2, SG3 and SG4 (p=0.041, p=0.37 and p=0.75 respectively), SG2 compared to SG3 (p=0.019) and SG4 (p=0.039), and SG3 compared to SG4 (p=0.33).

Table 7.8: Odds ratio of an unsafe response among all (correct and incorrect) responses by certainty and score group

<table>
<thead>
<tr>
<th>Score group</th>
<th>SG1 (n=20)</th>
<th>SG2 (n=159)</th>
<th>SG3 (n=133)</th>
<th>SG4 (n=18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low certainty</td>
<td>1.0</td>
<td>0.73</td>
<td>0.38</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>(0.49–1.11)</td>
<td>(0.24–0.58)</td>
<td>(0.08–0.43)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p=0.14</td>
<td>p&lt;0.0001</td>
<td>p &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>Moderate certainty</td>
<td>0.55</td>
<td>0.52</td>
<td>0.30</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>(0.32–0.97)</td>
<td>(0.34–0.77)</td>
<td>(0.06–0.25)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p=0.038</td>
<td>p=0.001</td>
<td>p &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>High certainty</td>
<td>0.90</td>
<td>0.37</td>
<td>0.32</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>(0.50–1.63)</td>
<td>(0.24–0.56)</td>
<td>(0.12–0.55)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p=0.73</td>
<td>p&lt;0.0001</td>
<td>p&lt;0.0001</td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio, (95% confidence interval)

7.4.6 Relationship between levels of certainty and the odds of an unsafe response among incorrect responses

When students gave an incorrect response, questions answered with high certainty were more likely to be unsafe than those answered with low certainty (OR 2.2; 95%CI, 1.7–2.8; p<0.0001) or with moderate certainty (OR 2.3; 95%CI, 1.8–2.8; p<0.0001).
7.4.7 Relationship between levels of certainty and the odds of an unsafe response among incorrect responses by ability group

The levels of certainty were significantly different between score groups (interaction of score group and level of certainty $F=3.4$, df=6446, $p=0.002$, Table 7.9). There were significant differences among levels of certainty for SG2 ($p=0.0007$), SG3 ($p<0.0001$) and SG4 ($p=0.014$), but not for SG1 ($p=0.13$). For SG2 and SG3, high certainty was significantly more likely to be associated with an unsafe response than moderate certainty (OR 1.8; 95%CI 1.3–2.4; $p=0.0001$; OR 3.6; 95%CI, 2.4–5.6 $p<0.0001$ respectively). For SG2, SG3 and SG4 scorers, high certainty was significantly more likely to be associated with an unsafe response than low certainty (OR 1.5, 95%CI 1.1–2.1, $p=0.006$; OR 4.3, 95%CI 2.8–6.6, $p<0.0001$; and OR 9.2, 95%CI, 2.0–42.0, $p=0.004$ respectively).

After adjusting for multiple comparisons, those students in SG2 responded differently in terms of unsafe, when incorrect, with increasing certainty compared to SG3 ($p=0.001$). SG1 were not significantly different from SG2 ($p=0.76$), SG3 ($p=0.027$) or SG4 ($p=0.038$). SG2 were not significantly different from SG4 ($p=0.043$). SG3 and SG4 were not significantly different ($p=0.35$). The ORs of an unsafe response for high versus low certainty were significantly higher for SG3 than SG2 ($p=0.0002$). The OR for high versus moderate certainty for SG3 (OR 3.6) was significantly higher than SG2 (OR 1.8) ($p=0.007$).
### Table 7.9: Odds ratio of an unsafe response among incorrect responses by certainty and score group

<table>
<thead>
<tr>
<th>Score group</th>
<th>SG1 n=20</th>
<th>SG2 n=159</th>
<th>SG3 n=133</th>
<th>SG4 n=18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low certainty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low certainty</td>
<td>1.0</td>
<td>0.9 (0.6–1.4)</td>
<td>0.7 (0.4–1.1)</td>
<td>0.4 (0.2–1.1)</td>
</tr>
<tr>
<td>Moderate certainty</td>
<td>0.7 (0.4–1.3)</td>
<td>0.8 (0.5–1.2)</td>
<td>0.8 (0.5–1.3)</td>
<td>1.2 (0.5–3.28)</td>
</tr>
<tr>
<td>High certainty</td>
<td>1.4 (0.7–2.8)</td>
<td>1.4 (0.8–2.3)</td>
<td>2.9 (1.6–5.1)</td>
<td>4.0 (1.0–15.6)</td>
</tr>
</tbody>
</table>

Odds ratio (95% confidence interval)

### 7.4.8 Relationship between levels of certainty and the odds of moderate or highly unsafe responses among all (correct and incorrect) responses

Highly unsafe responses have the potential to be the most significant with regards to adverse consequences. However, the number of highly unsafe responses chosen was only 30/6600. Therefore, to better analyse for those unsafe responses that were likely to have the greatest adverse consequences, those deemed moderate and highly unsafe were combined.

Students were less likely to give a moderately or highly unsafe response with increasing certainty. The OR for moderate vs. low certainty was 0.53 (95%CI, 0.36–0.77, p=0.0009), high vs. low 0.35 (95%CI, 0.23–0.53, p<0.0001) and high vs. moderate 0.66 (95%CI, 0.46–0.93, p=0.019).
7.4.9 Relationship between levels of certainty the odds of moderate or highly unsafe responses among all (correct and incorrect) responses by score group

The levels of certainty were not significantly different between score groups (interaction of score group and level of certainty F=1.4, df=6570, p=0.21, Table 7.10).

Table 7.10: Odds ratio of a moderately or highly unsafe (clinically significant) response among all (correct and incorrect) responses by certainty and score group

<table>
<thead>
<tr>
<th>Score group</th>
<th>SG1 n=20</th>
<th>SG2 n=159</th>
<th>SG3 n=133</th>
<th>SG4 n=18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low certainty</td>
<td>1.0 (0.30–1.25) p=0.18</td>
<td>0.61 (0.20–0.82) p=0.012</td>
<td>0.33 (0.06–0.33) p&lt;0.0001</td>
<td>0.12 (0.02–0.96) p=0.046</td>
</tr>
<tr>
<td>Moderate certainty</td>
<td>0.33 (0.10–1.08) p=0.067</td>
<td>0.40 (0.20–0.82) p=0.012</td>
<td>0.15 (0.06–0.33) p&lt;0.0001</td>
<td>0.06 (0.01–0.47) p=0.008</td>
</tr>
<tr>
<td>High certainty</td>
<td>0.79 (0.27–2.25) p=0.65</td>
<td>0.15 (0.06–0.37) p&lt;0.0001</td>
<td>0.14 (0.05–0.35) p&lt;0.0001</td>
<td>0.12 (0.01–0.94) p=0.044</td>
</tr>
</tbody>
</table>

Odds ratio (95% confidence interval)

7.4.10 Relationship between levels of certainty the odds of moderate or highly unsafe responses among incorrect responses

When students gave an incorrect response, there was no significant difference in the chance of a moderately or highly unsafe response between levels of certainty (F=2.3, df=2452, p=0.098). The OR for moderate certainty versus low was 0.70 (95%CI, 0.50–0.98, p=0.038), high vs. low 0.74 (95%CI, 0.47–1.16, p=0.19) and high vs. moderate 1.06 (95%CI, 0.67–1.66, p=0.82).
7.4.11 Relationship between levels of certainty the odds of moderate or highly unsafe responses among incorrect responses by ability group

The levels of certainty were not significantly different between the score groups (interaction of score group and level of certainty F=0.8, df=6446, p=0.54, (Table 7.11).

Table 7.11: Odds ratio of a moderately or highly unsafe (clinically significant) response among incorrect responses by certainty and score group

<table>
<thead>
<tr>
<th>Certainty Level</th>
<th>Scorer Group</th>
<th>SG1 n=20</th>
<th>SG2 n=159</th>
<th>SG3 n=133</th>
<th>SG4 n=18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low certainty</td>
<td></td>
<td>1.0</td>
<td>0.71</td>
<td>0.56</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.34–1.47)</td>
<td>(0.25–1.23)</td>
<td>(0.03–2.26)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p=0.36</td>
<td>p=0.15</td>
<td>p=0.23</td>
<td></td>
</tr>
<tr>
<td>Moderate certainty</td>
<td></td>
<td>0.39</td>
<td>0.56</td>
<td>0.32</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.12–1.32)</td>
<td>(0.27–1.16)</td>
<td>(0.04–3.08)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p=0.13</td>
<td>p=0.12</td>
<td>p=0.36</td>
<td></td>
</tr>
<tr>
<td>High certainty</td>
<td></td>
<td>1.01</td>
<td>0.38</td>
<td>0.52</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.34–2.98)</td>
<td>(0.16–0.92)</td>
<td>(0.07–4.86)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p=0.99</td>
<td>p=0.032</td>
<td>p=0.17</td>
<td>p=0.61</td>
</tr>
</tbody>
</table>

Odds ratio (95% confidence interval)
7.5 Discussion

Initially the results will be summarised in the context of the hypotheses and research questions. The results will be discussed related to feasibility of certainty-in and safeness-of responses, correctness for levels of confidence (certainty), and safeness for levels of confidence (certainty). Following this finding not related to the research questions will be discussed, specifically the internal consistency of response certainty, correctness and safeness. Lastly, the strengths and limitations of this study are discussed.

As a summary of the results for the hypotheses:

- Undertaking certainty-in and safeness-of MCQ responses as a format for an assessment used to inform high-stakes decisions being feasible was confirmed.

- Students of all ability levels, characterised by criterion, would demonstrate increasing correctness with certainty was confirmed.

- Unsafe incorrect responses would be held with increasing certainty by those with increasing levels of ability was confirmed.

As a summary of the results for the research questions:

1. Can certainty-in and safeness-of responses be included in an MCQ examination and used to inform high-stakes decisions?
   - Certainty in, and safety of, MCQ responses were used to inform high-stakes decisions.

2. What is the relationship between response certainty and odds of correctness?
   - Students had greater odds of correct responses as certainty increased.

3. What is the relationship between response certainty and the odds of potentially unsafe consequence responses?
   - Students had lower odds of incorrect and unsafe responses for moderate and high certainty.
   - When a student response was incorrect, the odds of such responses being unsafe were higher with high certainty.
4. How do these relationships of certainty with correctness and unsafeness vary with ability groupings, derived from the number correct on the assessment?

- All ability groupings, including the lower scoring groups, showed evidence of increasing odds of correctness with increasing certainty.
- When a student response was incorrect, the odds of such responses being unsafe were higher when such responses were also held with high certainty. This was seen in all groups except the lowest performers.

7.5.1 Feasibility of certainty-in and safeness-of responses

This study has shown that considering certainty-in and safeness-of responses can add to the information gained about, and for, medical students. The information gained regarding student ability from certainty in responses adds to that obtained from the number of correct responses alone. This information could be included in summative decision-making with a level of tolerance for deviations from optimal acceptable practice. A student may score above the pass threshold for number correct, and yet still be deemed not to have passed if they displayed sufficient unsafe responses that were held with high certainty.

The data gained from these question types were used to provide feedback with more information to students by including elements of metacognition, correctness for levels of certainty, and safeness for levels of certainty. In the assessment described in this study, the feedback in terms of correctness for levels of certainty and safeness for levels of certainty was presented specific to the question. In this way, by using these descriptors, and safeness, the context specific nature of certainty was recognised.

When using the certainty-in and safeness-of responses in the assessment of medical students, rules regarding tolerance, making allowances for responses that are incorrect, including unsafe, are required. This tolerance will allow for assessment to guide learning. For this assessment, these rules were developed by consensus amongst staff. No person or system is perfect. Deviations from optimal acceptable practice will occur, but only some lead to adverse outcomes and patient harm. This can range from no clinically-significant consequence to fatality. Systems and staff training are in place to minimise the frequency and consequences of decisions that are deviations from optimal acceptable practice. Zero tolerance of deviations...
from optimal acceptable practice, adverse outcomes and patient harm, is overly optimistic, but reducing the degree of harm from adverse events is highly desirable. Therefore, tolerance of deviations will vary dependant on the context and what other safeguards are in place. Even if no clinically significant adverse event or patient harm occurs, deviations should not be ignored, and should be considered as learning opportunities.

7.5.2 Correctness for levels of certainty

Within the cohort as a whole, and within the four performance groups, students had greater odds of correct responses as their levels of certainty increased. The lower scoring groups (SG1 and SG2) demonstrated evidence of increasing odds of correctness with increasing certainty. The lowest scoring group (SG1) had increasing odds of being correct between low and moderate certainty, and between low and high certainty, but there was no significant difference between moderate and high certainty. The below-standard group (SG2) demonstrated increasing odds of being correct with increasing certainty across all levels of certainty. In part, this confirms the findings from Study 1 and some recent literature, but is contrary to a general literature and healthcare literature which finds that those of lesser skill are also less aware of their underperformance. The reasons these results may differ from other literature may relate to the manner in which certainty was assessed. Unlike those with contrary results, the current study used certainty descriptors that refer to seeking assistance: this is more authentic to practice. Other researchers have used probabilistic predictions for certainty, such as estimating a probability that a response is correct, a format that may not be appropriate in all contexts. Instead, written descriptors for certainty may be preferable in some situations, such as in complex healthcare decisions. In other words, clinicians do not commonly ask if they are 30% likely to be correct or not. Instead, they ask themselves if they need extra information or help. The wording of certainty descriptors can alter responses, and therefore meaning. In the current study, certainty reflects self-assessment of performance done in-the-moment rather than retrospectively, and is the reflection-in-action authentic to daily practice, recognising the role of external cognitive support.
7.5.3 Safeness for levels of certainty

Within the cohort as a whole and the below-standard score (SG2) group, students had lower odds of incorrect and unsafe responses for moderate and high certainty. When a student response was incorrect, the odds of such responses being unsafe were higher in those answers that were also given with high certainty. This was seen in all groups except the lowest performers (SG1). If this were true in practice, then when clinicians happen to be incorrect but also believe they are correct and certain, then they are also more likely to be unsafe. This is clearly a situation that could compound circumstances towards patient harm.

The higher odds for unsafe responses when incorrect as certainty increases is consistent with Study 1. Replicating aspects of previous findings in different student groups increases the generalisability of the results. Previously, others reported that the number of unsafe responses is proportional to the numbers of incorrect responses, and therefore this additional metric adds nothing. However, all previous investigations have been retrospective, where the candidates were not given any scoring disincentive for responding unsafely.

Why might some students (SG2, SG3 and SG4) have a greater proportion of incorrect responses that are unsafe when they are highly certain? In addition to those factors already discussed in Study 1, students with higher ability levels (e.g., pass and excellent) may be more accustomed to take risks to increase their scores, and this may increase unsafe responses. This might lead to increasing unsafe responses with higher certainty.

One reassuring finding was that for those unsafe responses that are more likely to be clinically significant (moderate or highly unsafe), the odds of such responses amongst incorrect responses did not significantly vary with certainty. It is possible that such clinically significant responses were not detected because of the low numbers of such responses with higher certainty.

There were only four responses that were deemed incorrect and highly unsafe held with high certainty: one each from four students. These are rare events, and it would be important to recognise them if they occurred in practice, but they may have arisen due to keyboard error, rather than cognitive or metacognitive error. The problem of reliably assessing for rare events, such as unsafe responses, has been
raised and it remains an important issue. It is possible that multiple pieces of information gathered over a prolonged time, or a failure of a student’s decision-making behaviours to change even in the light of feedback, might add to more robust information.

7.5.4 Findings not related to research questions

The finding from Study 1 with the internal consistency of certainty being higher than that for correctness and safeness, was demonstrated in Study 2. The number of questions in these tests, being just 20, is a factor that potential limits these internal consistency results. Again this is an interesting result but extrapolation must be limited. These analyses will be reviewed in the discussion sections of subsequent studies and also related to all studies in this thesis in section 10.3.2.

7.5.5 Strengths of this Study

This study further reinforces some of the findings from Study 1, with an increase in the number of items and the number of students. In addition, this was in the setting of an assessment used to inform high-stakes pass-fail progression decisions. In moving away from a single score and looking at components of number correct for informing high-stakes pass-fail progression decisions, associated tolerances for deviations in correctness for certainty and unsafeness for certainty were required, and proved feasible.

7.5.6 Limitations of this study

The low scoring (SG1) and excellent groups (SG4) had relatively small numbers of students, which may have limited the ability to detect statistically significant differences within and between these groups, and resulted in wide confidence intervals for these comparisons. However, the below-standard (SG2) and above-standard (SG3) groups were of larger sizes, meaning that effects present in one of these and not the other is unlikely to be due to sampling.

The number of questions in the test was limited, and given that it was covering all the various aspects of the module, this led to the relatively low measures of internal consistency and high standard errors of measurement. To be considered as a stand-alone test with high-stakes decisions, an increase in question numbers will be required.
Another limitation of Study 2 relates to whether these measures truly reflect the degree of certainty or consideration of consequence for the students. There is no comparison with other measures or other assessments. Although the assessment did inform some summative decisions, it cannot be extrapolated that this would reflect current or future practice.

As for Study 1, are the unsafe responses truly unsafe? The panellists’ judgements showed kappa for these judgements was only fair. Although a weighted kappa may have been higher, for high-stakes assessment, such as stand-alone assessments, a larger number of reviewers for each incorrect response would be required, perhaps augmented by discourse.
8.1 Introduction

8.1.1 Building on prior studies

Building on Study 1, Study 2 and prior research, as outlined in the literature review, there are several aspects that need to be developed for Study 3. Some of these are to confirm previous findings in different contexts, some to investigate issues raised from Study 1 and Study 2, and some to explore areas highlighted from the literature review.

Both Study 1 and Study 2 were limited by a relatively small number of questions, and hence limited sampling of potential content. Given case specificity, whereby performance in one area of content or context does not confirm performance in another, it is important to assess across a range of contents and contexts. For Study 3, MCQ tests with a larger number of questions, covering more diverse content, needed to be investigated.

Both Study 1 and Study 2 report results on single tests. Although student feedback was included in both studies, it is not known if it was useful to students, or if it would alter their subsequent test performance or learning behaviour. For Study 3, two tests, with feedback on performance in between, needed to be investigated.

If MCQ tests incorporating certainty-in and safeness-of responses are assessing some cognitive functions required for practice, it would be expected that these functions would develop as medical students progressed through their programme of study. This being the case, it would be expected that there would be a degree of association between outcomes of MCQs incorporating certainty-in and safeness-of responses and other outcomes from other concurrent Otago Medical School (OMS) medical students’ assessments. It would not be expected that these associations would be very strong, as that would suggest that using certainty-in and safeness-of responses added no value to the concurrent assessments which do not include these parameters. For Study 3, investigation of test results with OMS assessment outcomes needed to be included.

As covered in the literature review, student demographics have been reported as being associated with variations in responses and outcomes in assessments. For Study 3, investigation of certainty-in and safeness-of response MCQs results with
demographics such as students’ self-reported gender and ethnicity needed to be included.

High certainty in unsafe responses would be potentially harmful to patients if those responses occurred in real practice. From Study 1 and Study 2, these kinds of episodes were rare and unpredictable. Such responses might be due to risk-taking behaviour, at least in the assessments. Assessment of risk-taking has been previously explored by others using the MCQ format. As described in a summary of the literature in section 3.5.7.2, a test was deliberately created with questions with no correct answer in order to ascertain if candidates will take risks in order to increase scores. Undergraduates, but not post-graduates, of lower ability were found to be more likely to take risks. For Study 3, questions designed to potentially identify risk-taking behaviours needed to be investigated.

The utility of the certainty-in and safeness-of response format to students would add to the evaluation. For Study 3, students’ perception of the format needed to be investigated.

8.1.2 Hypotheses and research questions

The hypotheses for Study 3 became:

- Undertaking certainty-in and safeness-of MCQ responses as a format for an MCQ test with a larger number of questions covering more diverse content would be feasible.

- Students of all ability levels, characterised by specified criteria, would demonstrate increasing correctness with certainty.

- Among students with higher ability, unsafe responses would be more likely as levels of certainty increase.

- Some, but not all, results for tests with certainty-in and safeness-of response MCQ questions would demonstrate a degree of association with OMS assessments results.

- Candidate gender and ethnicity would impact on certainty-in and safeness-of MCQ responses.

- Elapsed time and feedback and would lead to an improvement in response correctness and unsafeness for certainty.
• Student responses would demonstrate risk-taking patterns.
• Student perceptions on approach to, feedback from, and utility of certainty-in and safeness-of responses in MCQ format would be positive.

The research questions for Study 3 became:

1. Can certainty-in and safeness-of responses be included in an MCQ examination with a large number of questions?

2. What is the relationship between response certainty and correctness?

3. What is the relationship between response certainty and the potentially unsafe responses?

4. How do relationships of certainty with correctness and unsafeness vary with ability (number correct) groupings?

5. How do certainty-in and safeness-of response assessment results compare with other assessment results?

6. How do the results of certainty-in and safeness-of response assessments compare between gender and ethnicity groups?

7. How do certainty-in and safeness-of responses vary in questions modified to detect risk-taking behaviour?

8. How does time elapsed / feedback alter response patterns for certainty-in and safeness-of responses?

9. What do students perceive regarding their approach to, feedback from, and utility of, certainty-in and safeness-of response MCQs?
8.2 Methods

8.2.1 Design

To explore the research questions for Study 3, a variety of methods were used. They included the presentation of two research tests to students with provision of feedback between the tests. The results of Test 1 were compared to standard OMS assessment results, and analysed with respect to student gender and ethnicity. In addition, a subset of questions was developed to assess risk-taking behaviour. After Test 2, students were invited to be interviewed to discuss their perceptions of the tests’ formats.

8.2.2 Question pool

8.2.2.1 Question pool content, review and standard-setting process

To create a question bank for this experiment, 180 new questions were created by myself. Two questions were created for each of the relevant core presentations for the 2012 OMS MBChB degree course Curriculum Map (Appendix E).

The questions were developed to have 6-26 response options. The last option was the same for all questions, being, "I do not know and would ask someone else or consult a reference.”

A panel of five experts, including four hospital-based senior clinicians and a senior GP, all of whom were also University Staff members, reviewed content, suggested changes and developed the questions. Following this, the questions were finalised and the correct answers to each question confirmed.

The standard-setting process used was the Modified Angoff\textsuperscript{328}. This same process is used for standard-setting the OMS common component examinations. The five-member panel who standard-set the OMS Medicine Course examinations was used to standard-set the Test 1 and Test 2 questions.

The panel was asked to provide judgements against three standards, matched to the various major transitions within the course: “Minimum to exit Year 3, enter Year 4 of the course”; “Minimum to exit Year 5, enter Year 6 of the course”; and “Minimum to exit Year 6, enter PGY1”. Discussion by the panel members ensured that each had a
shared understanding of those standards. Each panellist independently judged the percentage of students demonstrating a performance for each of these standards in terms of how many would get the question correct. This was done using a scale of 0-100%, in 10% increments. The mean of all the judgements made by the panellists was used to create the estimated pass mark for each question.

A level of variance between panellists was expected. However, as the panellists were being asked to make judgements across multiple year levels, there may have been greater familiarity by individual panellists with some year levels than with others. If the judgment from each of the panellists of the standard for each year level is equivalent, then it is expected that variance in judgements will be the same for each year level. If the variances for the year levels were different would indicate a significant limitation in using the Modified Angoff method. If the variance is greater, then the accuracy of the judgements will be diminished. The F test was used to compare the variance in the five judgements across levels for each question. F tests assess the probability that the variance is not different. Subsequently, three F test analyses were made between year levels, exit Year 3 with exit Year 5, exit Year 5 with exit Year 6, and exit Year 3 with exit Year 6.

As expected, the mean of the judgements for the questions increased from Year 3 to 5, and 5 to 6. Only 14 of the 180 questions had F test analysis of <0.5 for all the three comparisons of levels, indicating that variances for these were likely different across all three. Of these 14, there was no particular question content area identified, and for 12 of the 14 the highest variance for the judgements for the exit Year 3 judgements.

8.2.3 Question classification for safeness process

Simultaneous with the Angoff standard-setting process, the same five clinician panel members reviewed the incorrect question options and made judgements regarding safeness. Discussion of safeness of responses considered the safety of the decisions if undertaken at the exit Year 5 point, which is entry into the Trainee Intern year. The responses were classified as no, low, moderate, or high unsafeness. Again the judgements were made independently. The range of options including ‘not unsafe’ reflects the range that will occur in practice. The median for these four possible category responses from the clinician panel members was taken as the safeness level. The median was used because the data are ordinal, thus enabling results to be
arranged simply on the scale from ‘not unsafe’ to ‘high unsafe’. If the median was between categories, the more unsafe one was used. This part of the standard-setting process involved the panel meeting for a day, and a further day of independent work.

Of the 2245 incorrect responses, 199 were classified as not unsafe, 793 low unsafeness, 955 moderate unsafeness, and 298 high unsafeness. Regarding the level of agreement between the five panellists, over the 180 questions and 2245 incorrect responses the Kappa was 0.11 using Stata IC 16.1 (StataCorp LLC, College Station, Texas, USA).

8.2.4 Question selection for use in the examinations

For Test 1, a random selection of 75 questions from the 180 question pool was selected for each student.

For Test 2 question selection, questions that any student had responded with high certainty to an unsafe response on Test 1 given to all, and the remainder of the questions were randomly taken from the remaining pool questions.

In order to explore risk-taking, previous researchers have deliberately removed the correct response from a question to see if candidates select incorrect responses. It did not seem appropriate to replicate this process and deliberately mislead students, so a variation was developed. Further question pools were developed from the 180-question pool. Each of the questions was reviewed with a view to developing two further versions: one having the correct response removed from the response list, and the other retaining it. However for both versions a free-text response option was added: “The correct answer is not present. The correct answer is: [Type text here]”.

For a few questions this did not work and therefore those questions were removed. These unsuitable questions had a scenario-question lead-in combination. Examples of this might include when a lead-in question related to the most important factor to consider in planning patient care; a free-text response might now have new and different correct responses such as ‘change in patient condition’ or ‘change in patient preferences’.
This process produced a set of questions for use in examinations which comprised three versions:

1. Standard question with no free-text box;
2. Correct option removed with the free-text box
3. Correct option present with the free-text box.

8.2.5 Subjects: candidates, recruitment, consequence/stakes

Volunteers were recruited via an invitation to participate email sent to all students in Years 2 to 5. An information sheet about the study was included with the invitation email. All who agreed to participate gave electronic consent.

The test results did not contribute to any progress decisions, and students were free to withdraw at any stage.

8.2.6 Test format and provision of information to candidates

The tests were delivered using a platform developed by an education technologist at the University of Otago. Each test was delivered at each of the four campuses, during three evenings (and therefore outside normal curriculum time).

Each test was invigilated in near examination conditions, with a single invigilator in the room and no resources or communication allowed. Students were advised to expect the tests to take 2½ hours, but no time constraints were imposed.

The descriptors for each of the question response options were updated from Study 2 and shown in Table 8.1. The front page of the test included a reminder to the students of the descriptors for levels of certainty.
Table 8.1: Certainty descriptors

<table>
<thead>
<tr>
<th>Level of certainty</th>
<th>None (Don’t know)</th>
<th>Low (Need direction)</th>
<th>Moderate (Need confirmation)</th>
<th>High (Action)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related to knowledge</td>
<td>You have no idea of correct response and therefore any response would be a guess.</td>
<td>You have no clear idea of correct response but have some knowledge on the subject. Any response would be based on limited information.</td>
<td>You have a reasonable idea of correct response on a basis of moderate knowledge on the subject. Any response would be based on sufficient information.</td>
<td>You are certain of correct response on a basis of detailed knowledge on the subject. Any response would not be a guess.</td>
</tr>
<tr>
<td>Related to practice</td>
<td>You would need to consult a colleague or references prior to considering any response.</td>
<td>You would need to consult a colleague or references but would be able to give a response first.</td>
<td>You would need to consult a colleague or references to confirm the correctness of the response.</td>
<td>You would have no need to consult a colleague or reference.</td>
</tr>
</tbody>
</table>
The instructions for candidates were as follows:

This examination consists of 75 questions. Each question is linked to a topic and drawn from a pool of 180 questions that align to the Faculty of Medicine Curriculum Map Presentations and Conditions.

Most of the questions consist of a scenario (such as the history, examination and investigation results related to a patient), a question task (such as “What is the most likely diagnosis?”) and a number of possible responses. The number of responses varies between questions.

One of the responses is always “I do not know and would ask someone else or consult a reference.”

Where appropriate another option will be “The correct answer is not present. The correct answer is ________”. This is a free-text response. To answer in the free-text box you need to click the button next to the free-text box, then write in that box.

For your chosen response to each question, apart from “I don’t know” you will indicate your certainty in that response.

For Test 1, a total of 75 questions were chosen at random for each student. Of these, 55 were standard format; 10 had a free-text box with removal of the correct answer; and 10 had a free-text box with the correct answer present.

8.2.7 Feedback to students

Feedback to the students was sent by email within two weeks of the last student attempting each test, and consisted of a table of individual question topics with correctness, certainty and safeness. No normative data were included in the feedback."
8.3  Analysis of test responses

8.3.1  Response and scoring analysis

Total numbers of responses by correctness, certainty and safety are reported.

As no two students sat the same Test 1 and no questions were sat by all students, it was not possible to undertake analysis for internal consistency using Cronbach’s $\alpha$ for correctness, certainty or safeness. This is because the Cronbach’s $\alpha$ analysis is across questions sat by all candidates.

Internal consistency is reported for Test 2 for the questions answered by all students. For the remaining other questions that were selected at random, internal consistency analysis using Cronbach’s $\alpha$ could not be applied.

8.3.2  Comparison between and among groups

Total responses were summed for correctness, certainty and safety. Proportions of responses for cohorts were calculated for group and response (correct, incorrect, unsafe, unsafe of incorrect) and compared using standard error or proportions and confidence intervals $^a$.

For the odds ratio (OR) analyses, differences between question types in the odds of a correct answer, or a ‘don’t know’ answer, were analysed with mixed model logistic regressions with a fixed term for question type, and random terms for student and the interaction of question and question type. The odds of giving a more unsafe answer were analysed with similar ordinal logistic regression models. The GLIMMIX procedure (SAS 9.4, SAS Institute Inc, North Carolina, USA) was used. Laplace’s method was used for the ordinal logistic regressions. For the ordinal logistic regressions among incorrect responses, the estimated G matrixes were not positive definite, and the random term for student was removed since they were estimated as zero.

8.3.3  Correctness and safeness by certainty, for ability groupings.

The ability groups were defined using the minimum standard scores from the Modified Angoff process, as being above level of minimum acceptable ability for stage of course or below level of minimum acceptable ability for stage of course.
The one variable used to analyse for difference in responses was:

- Above standard for year group.
- Below standard for year group.

The score equivalent to being above standard differed between years, but also differed between students for each year groups, as they sat different questions. These question groups have different associated pass marks. Given the different question pools, this also means that the odds of an unsafe response will vary between candidates.

### 8.3.4 Results by gender and ethnicity

All the results described for the study (certainty and safety) were analysed for difference based on gender and ethnicity. At enrolment, university students complete demographic information as part of the registration process.

This information included student gender:

- Male.
- Female.
- Not recorded/unknown.

This information included student ethnicity:

- European / Pākehā.
- Māori.
- Asian.
- Pacific.
- Middle Eastern / Latin American / African.
- Other.
- Not recorded/unknown.
8.3.5 Associations between study (certainty and safety) test results and concurrent OMS assessment results

In looking for associations between the study results (certainty and safety) and concurrent OMS assessment results, several different results were available for analysis.

As noted in section 4.4.6, there are several whole-cohort assessments that occur during the course, and the results for these will be used.

Concurrent assessment results include:

- For all students: RKT1 total score, RKT2 total score.
- For Year 2 students: OSCE score, OSPE score, SAQ score.
- For year 3 students: OSCE score, OSPE score, SAQ score.
- For Year 5 students: OSCE score, Written score.

The results used will be those that occurred during the year of the study test.

To look for associations among these results, Pearson’s correlation coefficient analysis was undertaken, with the statistical significance and strength reported. Given the multiple correlations, a Bonferroni correction (0.05/55) was undertaken, and $p < 0.001$ was used to demonstrate statistical significance.

8.3.6 Risk taking

Question format included standard questions, free-text box, correct response removed, and free-text box, correct response present. For those questions with the free-text box option, the students could elect to use the free-text box or the option list.

For each of the five possible question format and response selection combinations, total responses and proportions correct, unsafe, unsafe of incorrect, in total and for high certainty responses, and 'don’t know' responses were calculated.
Analysis was undertaken comparing the proportion for each of these responses in total for:

- Standard vs. non-standard formats.
- Standard vs. non-standard, correct option present.
- Standard vs. non-standard, correct option removed.
- Standard vs. non-standard, option selected.
- Standard vs. non-standard, free-text box selected.
- Non-standard format, correct option present vs. non-standard format, correct option removed.
- Non-standard format, option selected vs. non-standard format, free-text box selected.

Given the multiple comparisons, a Bonferroni correction (0.05/7) was undertaken, and p < 0.007 taken to indicate statistical significance.

8.3.7 Effect of time/feedback

For the students that sat both tests, cohort level responses for the number of responses by certainty, correctness and safety were totalled. From this, proportions for correctness, unsafeness and unsafeness of incorrect were calculated by level of certainty, and compared between the tests. A Bonferroni correction (0.05/24) was applied for multiple comparisons, and p<0.002 taken to indicate statistical significance.
8.4 Student interview process

Following Test 2, purposeful sampling was planned to invite those students who either had:

- Undertaken both Test 1 and Test 2.
- Selected any high certainty highly unsafe response on either Test 1 or Test 2.

The prospective participants were invited to an interview lasting up to an hour, either face-to-face or via videoconference (Skype). Times and venues were arranged to be mutually convenient, and a private room with no envisaged interruptions was chosen.

Interview participants gave written consent.

Following offering thanks for participation and checking the participant was ready to proceed, the interview commenced and was audio-recorded.

The interview questions were not fixed, but the following acted as an aide memoir:

- How did you approach the research assessment?
- How did you use the information you received from the research assessment?
- How do you approach faculty assessments?
- How did you use the information you received from faculty exams?
- What did you think of the format of certainty-in and safety-of questions:
  - To help your learning?
  - To be used to make progress decisions?

Recordings and notes were made, with recordings transcribed verbatim by an administrative assistant. The transcripts were read by myself whilst listening to the interview and re-read subsequently. Notes were then made on the transcripts regarding points being raised by the student. These points were written on paper notes then compared so consistent words were used for equivalent points. Then the points were grouped into themes. These themes were further developed with knowledge of the literature review and results of the studies within the thesis. This process aligns to the principles of thematic analysis and its use within medical education.
8.5 Results

A total of 89 students sat Test 1: 11 from Year 2, 17 from Year 3, 10 from Year 4, and 51 from Year 5. The low numbers from Years 2 to 4 does mean that these analyses might be underpowered to detect meaningful differences.

The total number of responses for all students is shown in Table 8.2.

Table 8.2: Total responses for all students

<table>
<thead>
<tr>
<th>Correctness</th>
<th>Certainty</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>–</td>
<td>802</td>
<td>959</td>
<td>490</td>
<td></td>
<td>2251</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Not unsafe</td>
<td>–</td>
<td>437</td>
<td>317</td>
<td>74</td>
<td>828</td>
</tr>
<tr>
<td></td>
<td>Low level unsafe</td>
<td>–</td>
<td>723</td>
<td>334</td>
<td>69</td>
<td>1126</td>
</tr>
<tr>
<td></td>
<td>Moderate level unsafe</td>
<td>–</td>
<td>552</td>
<td>321</td>
<td>56</td>
<td>929</td>
</tr>
<tr>
<td></td>
<td>High level unsafe</td>
<td>–</td>
<td>90</td>
<td>33</td>
<td>4</td>
<td>127</td>
</tr>
<tr>
<td>Don’t know</td>
<td>1235</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1235</td>
</tr>
<tr>
<td>Total</td>
<td>1235</td>
<td>2604</td>
<td>1964</td>
<td>693</td>
<td></td>
<td>6496</td>
</tr>
</tbody>
</table>

8.5.1 Correctness for levels of certainty

With increasing certainty there was an increase in proportion correct (Table 8.3). There were significant differences in the odds of a correct response between the levels of certainty (F=88.69 df=2,4679 p<0.0001). The OR of high vs. moderate certainty was 2.40 (95%CI, 1.91 - 3.01) p<0.0001.
### Table 8.3: Proportion correct and odds ratio of correct for levels of certainty for all

<table>
<thead>
<tr>
<th>Certainty</th>
<th>Percent correct</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>30.8% (802/2604)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>48.8% (959/1964)</td>
<td>1.94 (1.67 - 2.26)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High</td>
<td>70.7% (490/693)</td>
<td>4.66 (3.67 - 5.91)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

#### 8.5.2 Unsafeness for levels of certainty

With increasing certainty, the proportion unsafe decreased (Table 8.4). There were significant differences in the odds of an unsafe response between the levels of certainty (F=65.70 df=2,4679 p<0.0001). The OR of high vs. moderate certainty was 0.42 (95%CI, 0.32 - 0.55) p<0.0001.

### Table 8.4: Proportion unsafe and odds ratio of any unsafe for levels of certainty for all

<table>
<thead>
<tr>
<th>Certainty</th>
<th>Percent unsafe</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>52.4% (1365/2604)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>35.0% (688/1964)</td>
<td>0.55 (0.47 - 0.64)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High</td>
<td>18.6% (129/693)</td>
<td>0.23 (0.18 - 0.30)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

There were significant differences in the odds of a more unsafe response between the levels of certainty (F=59.38 df=2,4677 p<0.0001). The response was less likely to be more unsafe with increasing certainty. The OR of high vs. moderate certainty was 0.43 (95%CI, 0.33 - 0.55) p<0.0001 (Table 8.5).
Table 8.5: Proportion and odds by degree of unsafeness among all responses

<table>
<thead>
<tr>
<th>Certainty</th>
<th>Not unsafe / correct</th>
<th>Low unsafe</th>
<th>Moderate unsafe</th>
<th>High unsafe</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low (2604)</td>
<td>47.6% (1239)</td>
<td>27.8% (723)</td>
<td>21.2% (552)</td>
<td>3.5% (90)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Moderate (1964)</td>
<td>65.0% (1276)</td>
<td>17.0% (334)</td>
<td>16.3% (321)</td>
<td>1.7% (33)</td>
<td>0.61 (0.53 - 0.71)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High (693)</td>
<td>81.4% (564)</td>
<td>10.0% (69)</td>
<td>8.1% (56)</td>
<td>0.6% (4)</td>
<td>0.26 (0.20 - 0.34)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

The proportion of incorrect responses that are unsafe was high and was affected by level of certainty. Among incorrect responses there were significant differences in the odds of an unsafe response between the levels of certainty (F=3.49 df=2,2477 p=0.031) (Table 8.6). The OR of high vs. moderate certainty was 0.80 (95%CI, 0.47 - 1.35) p=0.40.

Table 8.6: Proportion and odds ratio of any unsafe within incorrect for levels of certainty for all

<table>
<thead>
<tr>
<th>Certainty</th>
<th>Percent unsafe</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>75.7% (1365/1802)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>68.5% (688/1005)</td>
<td>0.75 (0.58 - 0.97)</td>
<td>0.031</td>
</tr>
<tr>
<td>High</td>
<td>63.5% (129/693)</td>
<td>0.60 (0.36 - 1.00)</td>
<td>0.050</td>
</tr>
</tbody>
</table>

There were no significant differences in the odds of a more unsafe response between the levels of certainty (F=0.92 df=2,2475 p=0.40). The response was less likely to be more unsafe with increasing certainty. The OR of high vs. moderate certainty was 0.81 (95%CI, 0.56 - 1.17) p=0.26 (Table 8.7).
Table 8.7: Proportion and odds by degree of unsafeness among incorrect responses

<table>
<thead>
<tr>
<th>Certainty</th>
<th>Not unsafe</th>
<th>Low unsafe</th>
<th>Moderate unsafe</th>
<th>High unsafe</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low (1802)</td>
<td>24.3% (437)</td>
<td>40.1% (723)</td>
<td>30.6% (552)</td>
<td>5.0% (90)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Moderate (1005)</td>
<td>31.5% (317)</td>
<td>33.2% (334)</td>
<td>31.9% (321)</td>
<td>3.3% (33)</td>
<td>0.96 (0.79 - 1.16)</td>
<td>0.68</td>
</tr>
<tr>
<td>High (203)</td>
<td>36.5% (74)</td>
<td>34.0% (69)</td>
<td>27.6% (56)</td>
<td>2.0% (4)</td>
<td>0.78 (0.54 - 1.12)</td>
<td>0.18</td>
</tr>
</tbody>
</table>

8.5.3 Effect of ability on responses, Test 1

The year 5 group had the most respondents, but only two Year 5 students had a score above standard for exit Year 5, so no meaningful analysis of comparing responses by those above and below standard was possible. This low proportion above the pass threshold possibly relates to the nature of this assessment, being a research assessment with no stakes, and the timing being mid-year rather than end of year. The pass standard was that expected for an end-of-year high stakes assessment. In addition, the level of variance among the standard setting panel was higher than expected.

8.5.4 Effect of student ethnicity on responses, Test 1

The ethnicities recorded at registration for these students were European/Pākehā-only, 63 (71%); Asian-only, 18 (20%), European/Pākehā and Māori, 2 (2%), European/Pākehā and Pacific Peoples, 2 (2%), and 1 (1%) each were Middle Eastern/Latin American/African/European, Pacific Peoples and European/Pākehā, and Pacific Peoples. The lack of numbers of students apart from those who self-identified as European/Pākehā-only or Asian-only meant that no meaningful analysis of comparing responses across the different student ethnicities was possible.
8.5.5 Effect of student gender on responses, Test 1

The Test 1 respondents included 53 females, 32 males and 4 with no gender recorded. There was no significant difference between genders in terms of proportion correct (F=0.22 DF=1,5670 p=0.64) (Table 8.8).

Table 8.8: Proportion and odds ratio correct by gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Percent correct</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>34.1% (1323/3884)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>36.1% (859/2377)</td>
<td>1.08 (0.77 - 1.52)</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Excluding 'don’t know' responses, there was no significant difference between genders in terms of proportion correct (F=1.53 DF=1,4489 p=0.22) (Table 8.9).

Table 8.9: Proportion and odds ratio correct excluding ‘don’t know’ responses by gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Percent correct</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>41.8% (1323/3167)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>45.3% (859/1897)</td>
<td>1.18 (0.91 - 1.54)</td>
<td>0.22</td>
</tr>
</tbody>
</table>

There were significant difference in the odds of being more certain between genders (F=6.40 df=1,4488 p=0.011) (Table 8.10).

Table 8.10: Proportion and odds ratio of certainty of responses by gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (3167)</td>
<td>53.0% (1680)</td>
<td>36.9% (1169)</td>
<td>10.0% (318)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Male (1897)</td>
<td>42.4% (804)</td>
<td>38.3% (727)</td>
<td>19.3% (366)</td>
<td>2.00 (1.17 - 3.43)</td>
<td>0.011</td>
</tr>
</tbody>
</table>
There was no significant difference between genders in the differences in the odds of a correct response between the levels of certainty (F=0.79 df=2,4485 p=0.45, the certainty x gender interaction) (Table 8.11).

Table 8.11: Proportion and odds ratio of correct for levels of certainty comparing by gender

| Certainty | Female | | | Male | | |
|-----------|--------|----------|----------|--------|----------|
|           | Correct | OR (95% CI) | p | Correct | OR (95% CI) | p |
| Low       | 31.2% (524/1680) | 1.00 | | 30.2% (243/804) | 1.00 | |
| Moderate  | 49.6% (580/1169) | 1.94 (1.61 - 2.37) | <0.0001 | 48.0% (349/727) | 1.99 (1.54 - 2.56) | <0.0001 |
| High      | 68.9% (219/318) | 4.18 (3.01 - 5.82) | <0.0001 | 73.0% (267/366) | 5.58 (3.94 - 7.90) | <0.0001 |

There was a significant difference between genders in the odds of any unsafe response (F=4.28 DF=1,5670 p=0.039) with males were less likely to give an unsafe response (Table 8.12).

Table 8.12: Proportion and odds ratio for any unsafe response by gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Any unsafe</th>
<th>OR (95% CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>35.0% (1361/3884)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>30.4% (722/2377)</td>
<td>0.76 (0.59 – 0.99)</td>
<td>0.039</td>
</tr>
</tbody>
</table>

There were significant differences in the odds of a more unsafe response between genders (F=5.01 df=1,5668 p=0.025) (Table 8.13) with males making fewer unsafe responses at all levels of unsafeness.
Table 8.13: Proportion and odds ratio of unsafe responses by gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Not unsafe</th>
<th>Low unsafe</th>
<th>Moderate unsafe</th>
<th>High unsafe</th>
<th>OR (95% CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (3884)</td>
<td>65.0% (2523)</td>
<td>17.7% (689)</td>
<td>15.1% (588)</td>
<td>2.2% (84)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Male (2377)</td>
<td>69.6% (1655)</td>
<td>16.2% (385)</td>
<td>12.5% (297)</td>
<td>1.7% (40)</td>
<td>0.76 (0.59 - 0.97)</td>
<td>0.025</td>
</tr>
</tbody>
</table>

There was no significant difference between genders in the odds of a more unsafe response among incorrect responses (F=1.27 df=1,2358 p=0.26) (Table 8.14).  

Table 8.14: Proportion and odds ratio of unsafe responses within incorrect responses (excluding correct and 'don’t know' responses)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Not unsafe</th>
<th>Low unsafe</th>
<th>Moderate unsafe</th>
<th>High unsafe</th>
<th>OR (95% CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (1844)</td>
<td>26.2% (483)</td>
<td>37.4% (689)</td>
<td>31.9% (588)</td>
<td>4.6% (84)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Male (1038)</td>
<td>30.4% (316)</td>
<td>37.1% (385)</td>
<td>28.6% (297)</td>
<td>3.9% (40)</td>
<td>0.90 (0.76 - 1.08)</td>
<td>0.26</td>
</tr>
</tbody>
</table>

There was no significant difference between genders in terms of any unsafe response within incorrect responses (F=3.06 DF=1,2360 p=0.081) (Table 8.15).

Table 8.15: Proportion and odds ratio of any unsafe response within incorrect responses

<table>
<thead>
<tr>
<th>Gender</th>
<th>Any unsafe</th>
<th>OR (95% CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>73.8% (1361/1844)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>69.6% (722/1038)</td>
<td>0.80 (0.62 - 1.03)</td>
<td>0.081</td>
</tr>
</tbody>
</table>
There was no significant difference between genders in the differences in the odds of any unsafe response between the levels of certainty (F=0.04 df=2,4485 p=0.96, the certainty - gender interaction) (Table 8.16).

Table 8.16: Proportion and odds ratio of any unsafe response for levels of certainty comparing by gender

<table>
<thead>
<tr>
<th>Certainty</th>
<th>Unsafe (95%CI)</th>
<th>p</th>
<th>Unsafe (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>53.2% (894/1680)</td>
<td>1.00</td>
<td>50.6% (407/804)</td>
<td>1.00</td>
</tr>
<tr>
<td>Moderate</td>
<td>34.9% (408/1169)</td>
<td>0.53 (0.43 - 0.65)</td>
<td>&lt;0.0001</td>
<td>34.1% (248/727)</td>
</tr>
<tr>
<td>High</td>
<td>18.6% (59/318)</td>
<td>0.23 (0.16 - 0.33)</td>
<td>&lt;0.0001</td>
<td>18.3% (67/366)</td>
</tr>
</tbody>
</table>

Among incorrect responses (excluding correct and ‘don’t know’ responses) there was no significant difference between genders in the differences in the odds of any unsafe response between the levels of certainty (F=0.16 df=2,2356 p=0.85 the certainty x gender interaction) (Table 8.17).
Table 8.17: Proportion and odds ratio of any unsafe response within incorrect responses for levels of certainty comparing by gender

<table>
<thead>
<tr>
<th>Certainty</th>
<th>Female</th>
<th></th>
<th></th>
<th>Male</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unsafe</td>
<td>OR (95%CI)</td>
<td>p</td>
<td>Unsafe</td>
<td>OR (95%CI)</td>
</tr>
<tr>
<td>Low</td>
<td>77.3% (894/1156)</td>
<td>1.00</td>
<td>72.5% (407/561)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>69.3% (408/589)</td>
<td>0.71 (0.50 - 0.99)</td>
<td>0.045</td>
<td>65.6% (248/378)</td>
<td>0.83 (0.53 - 1.28)</td>
</tr>
<tr>
<td>High</td>
<td>59.6% (59/99)</td>
<td>0.59 (0.28 - 1.21)</td>
<td>0.15</td>
<td>67.7% (67/99)</td>
<td>0.62 (0.29 - 1.31)</td>
</tr>
</tbody>
</table>

8.5.6 Comparisons with standard OMS assessments

8.5.6.1 Retained Knowledge Tests

The Retained Knowledge Test (RKT) is sat by all students in Years 2 to 5. Therefore all the students who sat the research tests would have sat the same two RKTs in the year of the Study 3 tests. Correlations for total and proportions of responses varied (Table 8.18). The total correct in the two RKTs and study Test 1 correlated moderately strongly with each other ($r=0.55$ for both, $p<0.001$). There was no correlation between total unsafe and either of the RKT’s total correct.

The number correct in RKT1 correlated fairly with the proportion of responses correct with low certainty on Test 1 ($r=0.10$, $p<0.001$), and moderate certainty ($r=0.10$, $p<0.001$), but did not correlate with the proportion correct with high certainty. The number correct in RKT2 correlated poorly with the proportion of responses correct with moderate certainty ($r=0.08$, $p<0.001$), but did not correlate with the proportion correct with low or high certainty.

The number correct in RKT1 inversely correlated fairly with the proportion of unsafe responses with low certainty on Test 1 ($r=-0.11$, $p<0.001$), but not with the proportion unsafe responses with moderate nor high certainty. The number correct in RKT2 did not correlate with the proportion of unsafe responses for any of the
levels of certainty. None of the analyses for the number correct in RKT1 and RKT2 and the proportion of incorrect responses was significant.

8.5.6.2 Year 2 RKT and CCE

There were no significant correlations among the RKT1 score, RKT2 score, SAQ score, OSCE score, and OSPE score, and the Test 1 total correct and total unsafe, nor for the proportion correct, unsafe, or unsafe responses of incorrect responses with any level of certainty (Table 8.19).

8.5.6.3 Year 3 RKT and CCE

There were no significant correlations among the RKT1 score, RKT2 score, SAQ score, OSCE score, and OSPE score, and the Test 1 total correct and total unsafe, nor for the proportion correct, unsafe, or unsafe responses of incorrect responses with any level of certainty (Table 8.20).

8.5.6.4 Year 4 RKT

There were no significant correlations among the RKT1 score, RKT2 score, and the Test 1 total correct and total unsafe, nor for the proportion correct, unsafe, or unsafe responses of incorrect responses with any level of certainty (Table 8.21).

8.5.6.5 Year 5 RKT and CCE

The RKT1, RKT2, and written CCE scores all correlated with the Test 1 total correct ($r=0.28, r=0.26, r=0.17$ respectively, all p<0.001). The written CCE score inversely correlated with the number of unsafe responses in Test 1 ($r=-0.17$, p<0.001) (Table 8.22).
<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 1</th>
<th>Proportion correct for level of certainty</th>
<th>Proportion unsafe for level of certainty</th>
<th>Proportion unsafe of incorrect for level of certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total correct</td>
<td>Total unsafe</td>
<td>Low</td>
<td>Mod.</td>
<td>High</td>
</tr>
<tr>
<td>RKT1</td>
<td>0.55*</td>
<td>&lt;0.01</td>
<td><strong>0.10</strong>*</td>
<td><strong>0.10</strong>*</td>
<td><strong>0.11</strong>*</td>
</tr>
<tr>
<td>RKT2</td>
<td>0.55*</td>
<td>&lt;0.01</td>
<td>0.06</td>
<td><strong>0.08</strong>*</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*p<0.001, Pearson r² coefficient

Table 8.18: Correlation for standard assessments results among Test 1 total correct, total unsafe, and proportions of responses for certainty
Table 8.19: Correlation for standard assessment results among Test 1 total correct, total unsafe, and proportions of responses for certainty for Year 2 students

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 1</th>
<th>Proportion correct for level of certainty</th>
<th>Proportion unsafe for level of certainty</th>
<th>Proportion unsafe of incorrect for level of certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total correct</td>
<td>Total unsafe</td>
<td>Low</td>
<td>Mod.</td>
<td>High</td>
</tr>
<tr>
<td>RKT1</td>
<td>0.49</td>
<td>0.01</td>
<td>0.07</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>RKT2</td>
<td>0.10</td>
<td>0.01</td>
<td>0.01</td>
<td>&lt;0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>SAQ</td>
<td>0.53</td>
<td>0.01</td>
<td>0.05</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>OSCE</td>
<td>0.03</td>
<td>0.09</td>
<td>0.03</td>
<td>0.01</td>
<td>0.32</td>
</tr>
<tr>
<td>OSPE</td>
<td>0.06</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>0.17</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Pearson r² coefficient
Table 8.20: Correlation for standard assessments among Test 1 total correct, total unsafe, and proportions of responses for Year 3 students

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 1</th>
<th>Proportion correct for level of certainty</th>
<th>Proportion unsafe for level of certainty</th>
<th>Proportion unsafe of incorrect for level of certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total correct</td>
<td>Total unsafe</td>
<td>Low</td>
<td>Mod.</td>
<td>High</td>
</tr>
<tr>
<td>RKT1</td>
<td>0.02</td>
<td>0.09</td>
<td>0.41</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>RKT2</td>
<td>0.00</td>
<td>0.08</td>
<td>0.36</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>SAQ</td>
<td>0.00</td>
<td>0.05</td>
<td>0.28</td>
<td>&lt;0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>OSCE</td>
<td>0.00</td>
<td>0.19</td>
<td>0.31</td>
<td>0.22</td>
<td>0.03</td>
</tr>
<tr>
<td>OSPE</td>
<td>0.01</td>
<td>0.02</td>
<td>0.12</td>
<td>0.01</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Pearson r² coefficient
Table 8.21: Correlation for standard assessments among Test 1 total correct, total unsafe, and proportions of responses for Year 4 students

<table>
<thead>
<tr>
<th></th>
<th>Test 1 correct</th>
<th>Test 1 unsafe</th>
<th>Proportion correct for level of certainty</th>
<th>Proportion unsafe for level of certainty</th>
<th>Proportion unsafe of incorrect for level of certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Total</td>
<td>Low</td>
<td>Mod.</td>
<td>High</td>
</tr>
<tr>
<td>RKT1</td>
<td>0.26</td>
<td>&lt;0.01</td>
<td>0.01</td>
<td>0.14</td>
<td>0.01</td>
</tr>
<tr>
<td>RKT2</td>
<td>0.02</td>
<td>0.11</td>
<td>&lt;0.01</td>
<td>0.32</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Pearson $r^2$ coefficient
Table 8.22: Correlation for standardised assessments among Test 1 total correct, total unsafe, and proportions of responses for Year 5 students

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Test 1</th>
<th>Proportion correct for level of certainty</th>
<th>Proportion unsafe for level of certainty</th>
<th>Proportion unsafe of incorrect for level of certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total correct</td>
<td>Total unsafe</td>
<td>Low</td>
<td>Mod.</td>
<td>High</td>
</tr>
<tr>
<td>RKT1</td>
<td>0.28*</td>
<td>0.05</td>
<td>0.01</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>RKT2</td>
<td>0.26*</td>
<td>0.09*</td>
<td>&lt;0.01</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Written</td>
<td>0.18*</td>
<td>0.17*</td>
<td>0.03</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>OSCE</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
</tr>
</tbody>
</table>

*p<0.001, Pearson r coefficient
Table 8.23: Correlation among Test 1 total correct, total unsafe, and proportions of responses

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th>Proportion correct for level of certainty</th>
<th>Proportion unsafe for level of certainty</th>
<th>Proportion unsafe of incorrect for level of certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total Unsafe</td>
<td>Low</td>
<td>Mod</td>
<td>High</td>
</tr>
<tr>
<td>Test 1</td>
<td>Total correct</td>
<td>&lt;0.01</td>
<td>0.15*</td>
<td>0.15*</td>
</tr>
<tr>
<td>Test 1</td>
<td>Total unsafe</td>
<td>0.18*</td>
<td>0.10*</td>
<td>0.08</td>
</tr>
<tr>
<td>Proportion correct for level of certainty</td>
<td>Low</td>
<td>0.15*</td>
<td>0.02</td>
<td>0.69*</td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>0.08</td>
<td>0.14*</td>
<td>0.46*</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>0.05</td>
<td>0.11*</td>
<td>0.61*</td>
</tr>
<tr>
<td>Proportion unsafe for level of certainty</td>
<td>Low</td>
<td>0.14*</td>
<td>0.03</td>
<td>0.46*</td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>0.10*</td>
<td>0.04</td>
<td>0.42*</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>0.04</td>
<td>0.03</td>
<td>0.41*</td>
</tr>
<tr>
<td>Proportion unsafe of incorrect for level of certainty</td>
<td>Low</td>
<td>0.04</td>
<td>0.03</td>
<td>0.41*</td>
</tr>
<tr>
<td></td>
<td>Mod</td>
<td>&lt;0.01</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>

*p<0.001, Pearson r² coefficient
8.5.6.6 Interpretation of correlational analysis with standard OMS assessments

Not surprisingly, given that the format was identical and the content similar, the RKT1 score correlated very strongly with the RKT2 score ($r=0.79$, $p<0.001$). For each individual year this correlation was either very strong or moderately strong (Y2 $r=0.69$, $p<0.001$; Y3 $r=0.59$, $p<0.001$; Y4 $r=0.62$, $p<0.001$; Y5 $r=0.52$, $p<0.001$).

For the Year 3 CCE, there were significant correlations, being very strong for the SAQ and OSPE ($r=0.71$, $p<0.001$), and moderately strong for SAQ with RKT1 ($r=0.45$, $p<0.001$) and RKT2 ($r=0.50$, $p<0.001$). The SAQ and OSPE all cover underpinning and applied knowledge, although the formats and stakes of the assessments do vary, so such correlations are not surprising.

For the Year 5 CCE, there were fair correlations between the written score and the RKT1 and RKT2 scores ($r=0.26$ and $r=0.24$ respectively, $p<0.001$) and also between the written score and OSCE ($r=0.19$, $p<0.001$). Similar to Year 3, the RKT and CCE written assess similar content, and for part of the written exam, a similar format (MCQ). The CCE written and OSCE are completed within a few days of each other, and are likely to represent peak prepared performance, and although the content and format are different, a fair correlation between the two is not surprising.

8.5.6.7 Interpretation of correlational analysis of outcomes of the Test 1

The results of the analyses are shown in Table 8.23.

Unsurprisingly, the total correct correlated with the proportion correct for level of certainty, although this was only a fair correlation, and for low and moderate certainty ($r=0.15$, $p<0.001$ for both). The total correct was also inversely correlated, with fair strength, with the proportion unsafe with low certainty ($r=0.14$, $p<0.001$), and inversely correlated with poor strength with the proportion of unsafe incorrect responses with low certainty ($r=0.08$, $p<0.001$). These correlations may have been affected by the random selection of questions and that different students received different questions for study Test 1.

Likewise, the total unsafe correlated with the proportion unsafe for each level of certainty, being moderately strong for moderate certainty ($r=0.42$, $p<0.001$) and fairly correlated for low and high certainty ($r=0.22$ and $r=0.09$ respectively, $p<0.001$
The total unsafe correlated with the proportion of unsafe incorrect responses for each level of certainty, being fair for moderate certainty ($r=0.26$, $p<0.001$), and poor for low certainty ($r=0.08$, $p<0.001$). The total unsafe also inversely correlated with fair strength with the proportion correct for low and moderate certainty ($r=0.18$ and $r=0.10$ respectively, both $p<0.001$).

Taken together, these correlations imply that the proportion correct or unsafe for any level of certainty relates to the total correct and unsafe, and also that the proportion unsafe is a function of the proportion incorrect. However, the levels of correlation are not very strong, and the lower levels of correlation, where any existed for high certainty response, may imply that these responses are different. Given that high certainty responses are those equivalent to ‘action without the need to check’ in clinical practice, it might be expected to have a very high proportion correct with very few unsafe responses, irrespective of overall levels of correct. However, for some respondents the number of high certainty responses was low, and therefore the responses are potentially less reliable.

Within each level of certainty there were inverse correlations between correct and unsafe responses, and positive correlations with unsafe and unsafe incorrect responses. There were no correlations between the proportion correct and the proportion of unsafe incorrect responses. The inverse correlations for correct and unsafe were very strong for low certainty ($r=0.69$, $p<0.001$), and moderately strong for moderate certainty and high certainty ($r=0.46$ and $r=0.61$ respectively, $p<0.001$ for both). The correlations between unsafe and unsafe of incorrect were all moderately strong (low certainty $r=0.46$, moderate $r=0.43$, high $r=0.41$, all $p<0.001$).

There were also significant correlations at a fair strength, between adjacent levels of low and moderate certainty for correct ($r=0.15$, $p<0.001$), and for unsafe ($r=0.30$, $p<0.001$), and for moderate and high certainty for unsafe ($r=0.10$, $p<0.001$).

Lastly, there were significant inverse correlations of a fair strength between the proportion correct with moderate certainty and the proportion unsafe with low certainty ($r=0.14$, $p<0.001$), between the proportion correct with low certainty and the proportion unsafe with moderate certainty ($r=0.14$, $p<0.001$), and between the proportion correct with high certainty and the proportion unsafe with moderate certainty ($r=0.11$, $p<0.001$).
The fact that there were correlations between the adjacent levels of certainty, but not between low and high certainty suggests that there could be an ability to be increasingly correct and decreasingly unsafe for levels of certainty. The high certainty responses were slightly different, which could be explained by a level of correctness and unsafeness which varies less across the range of total correct, total unsafe, proportions correct or unsafe for low or moderate certainty.

8.5.7 Risk-taking

The total responses for questions of different formats and responses are reported in the following tables with analysis presented subsequently. These include: standard MCQ format (Table 8.24); free-text box present, correct response present, an option selected (Table 8.25); free-text box present, correct response present, free-text box selected (Table 8.28).

*Table 8.24: Standard format*

<table>
<thead>
<tr>
<th></th>
<th>Certainty</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>706</td>
<td>838</td>
<td>428</td>
<td>1972</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Not unsafe</td>
<td>-</td>
<td>364</td>
<td>262</td>
<td>62</td>
<td>688</td>
</tr>
<tr>
<td></td>
<td>Low level unsafe</td>
<td>-</td>
<td>596</td>
<td>277</td>
<td>61</td>
<td>934</td>
</tr>
<tr>
<td></td>
<td>Moderate level unsafe</td>
<td>-</td>
<td>448</td>
<td>257</td>
<td>42</td>
<td>747</td>
</tr>
<tr>
<td></td>
<td>High level unsafe</td>
<td>-</td>
<td>71</td>
<td>24</td>
<td>3</td>
<td>98</td>
</tr>
<tr>
<td>Don’t know</td>
<td></td>
<td>1016</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1016</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1016</td>
<td>2185</td>
<td>1658</td>
<td>596</td>
<td>5455</td>
</tr>
</tbody>
</table>
Table 8.25: Free-text box present, with correct response present, option selected

<table>
<thead>
<tr>
<th></th>
<th>Certainty</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td></td>
<td>72</td>
<td>71</td>
<td>36</td>
<td>179</td>
</tr>
<tr>
<td>Not unsafe</td>
<td></td>
<td>28</td>
<td>18</td>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>Low level unsafe</td>
<td></td>
<td>46</td>
<td>20</td>
<td>4</td>
<td>70</td>
</tr>
<tr>
<td>Moderate level unsafe</td>
<td></td>
<td>41</td>
<td>32</td>
<td>9</td>
<td>82</td>
</tr>
<tr>
<td>High level unsafe</td>
<td></td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Don’t know</td>
<td>103</td>
<td></td>
<td></td>
<td></td>
<td>103</td>
</tr>
<tr>
<td>Total</td>
<td>103</td>
<td>193</td>
<td>144</td>
<td>52</td>
<td>492</td>
</tr>
</tbody>
</table>

Table 8.26: Free-text box present, with correct response present, free-text box selected

<table>
<thead>
<tr>
<th></th>
<th>Certainty</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>Low</td>
<td>Moderate</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Not unsafe</td>
<td></td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Low level unsafe</td>
<td></td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Moderate level unsafe</td>
<td></td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>High level unsafe</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>13</td>
<td>11</td>
<td>2</td>
<td>26</td>
</tr>
</tbody>
</table>
From these data, the proportions of responses are calculated (Table 8.29).
Table 8.29: Proportion of cohort response by format/response

<table>
<thead>
<tr>
<th></th>
<th>Proportion correct</th>
<th>Proportion unsafe</th>
<th>Proportion of incorrect unsafe</th>
<th>Proportion 'don't know'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard (no free-text box)</td>
<td>0.36</td>
<td>0.33</td>
<td>0.72</td>
<td>0.19</td>
</tr>
<tr>
<td>Correct option present, options used</td>
<td>0.36</td>
<td>0.33</td>
<td>0.77</td>
<td>0.21</td>
</tr>
<tr>
<td>Correct option present, text used</td>
<td>0.19</td>
<td>0.58</td>
<td>0.47</td>
<td>0.00</td>
</tr>
<tr>
<td>Correct option removed, options used</td>
<td>0.00</td>
<td>0.52</td>
<td>0.73</td>
<td>0.29</td>
</tr>
<tr>
<td>Correct option removed, text used</td>
<td>0.81</td>
<td>0.12</td>
<td>0.64</td>
<td>0.00</td>
</tr>
<tr>
<td>All non-standard</td>
<td>0.27</td>
<td>0.39</td>
<td>0.73</td>
<td>0.21</td>
</tr>
<tr>
<td>Non-standard, all correct option present</td>
<td>0.36</td>
<td>0.34</td>
<td>0.73</td>
<td>0.20</td>
</tr>
<tr>
<td>Non-standard, correct option removed</td>
<td>0.18</td>
<td>0.43</td>
<td>0.72</td>
<td>0.22</td>
</tr>
<tr>
<td>Non-standard, option selected</td>
<td>0.20</td>
<td>0.42</td>
<td>0.75</td>
<td>0.24</td>
</tr>
<tr>
<td>Non-standard, text used</td>
<td>0.70</td>
<td>0.20</td>
<td>0.54</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The high certainty responses are those with the greatest potential for issues, as they represent action without the need for checking. It was reassuring that when the free-text box was used with high certainty, the candidates were correct 96% of the time, with no unsafe responses. When the correct option was removed and the candidate chose the free-text box, they were correct 100% of the time. However, they did not always choose to use the free-text box, as when the correct response was removed they were only correct 58% of the time (Table 8.30). This difference when responses
are chosen from the option list responses versus free-test responses is discussed subsequently.

Table 8.30: Proportions of cohort responses with high certainty by format/response outcome

<table>
<thead>
<tr>
<th></th>
<th>Proportion correct</th>
<th>Proportion unsafe</th>
<th>Proportion of incorrect unsafe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard (no free-text box)</td>
<td>0.71</td>
<td>0.08</td>
<td>0.42</td>
</tr>
<tr>
<td>Correct option present, options used</td>
<td>0.69</td>
<td>0.27</td>
<td>0.88</td>
</tr>
<tr>
<td>Correct option present, text used</td>
<td>0.50</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Correct option removed, options used</td>
<td>0.0</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Correct option removed, text used</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>All non-standard</td>
<td>0.64</td>
<td>0.26</td>
<td>0.66</td>
</tr>
<tr>
<td>Non-standard, all correct option present</td>
<td>0.69</td>
<td>0.26</td>
<td>0.82</td>
</tr>
<tr>
<td>Non-standard, correct option removed</td>
<td>0.58</td>
<td>0.21</td>
<td>0.50</td>
</tr>
<tr>
<td>Non-standard, option selected</td>
<td>0.51</td>
<td>0.33</td>
<td>0.68</td>
</tr>
<tr>
<td>Non-standard, text used</td>
<td>0.96</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Further analysis of cohort response proportions across the formats was undertaken. The proportion correct was higher, and the proportion unsafe lower, for the standard format compared with the non-standard formats (Table 8.31), which is consistent with responses from the option-list being different those free-text.

Table 8.31: Comparing responses by standard vs. all non-standard formats
When the correct option was present, the pattern was the same, with the proportion correct higher and the proportion unsafe lower for the standard format compared with the non-standard formats (Table 8.32).

Table 8.32: Standard vs. non-standard format, correct options present

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>All non-standard formats, correct option present</th>
<th>Difference (99.3%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0.36</td>
<td>0.18</td>
<td>0.18 (0.11 – 0.25)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>0.33</td>
<td>0.43</td>
<td>-0.11 (-0.18 - -0.04)</td>
</tr>
<tr>
<td>Unsafe of incorrect</td>
<td>0.72</td>
<td>0.72</td>
<td>0.00 (-0.10 – 0.09)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.19</td>
<td>0.22</td>
<td>-0.04 (-0.09 – 0.02)</td>
</tr>
</tbody>
</table>

p<0.007

When the correct option was removed, the proportion correct was significantly reduced, and the proportion unsafe increased, as shown in Table 8.33.

261
Table 8.33: Standard vs. non-standard format, correct option removed

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Non-standard, correct option removed</th>
<th>Difference (99.3%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0.36</td>
<td>0.18</td>
<td>0.18 (0.11 – 0.25)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>0.33</td>
<td>0.43</td>
<td>-0.11 (-0.18 - -0.04)</td>
</tr>
<tr>
<td>Unsafe of incorrect</td>
<td>0.72</td>
<td>0.72</td>
<td>0.00 (-0.10 – 0.09)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.19</td>
<td>0.22</td>
<td>-0.04 (-0.09 – 0.02)</td>
</tr>
</tbody>
</table>

p<0.007

When an option was selected on the non-standard format, although the DKO was used more, the response was a lower proportion correct, and higher proportion unsafe (Table 8.34). This could be due to a lack of knowledge, a lack of self-monitoring or a risk-taking strategy. For the non-standard format with an option selected with high certainty, the proportion correct was only 0.51.

Table 8.34: Standard vs. non-standard format, option selected

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Non-standard, option selected</th>
<th>Difference (99.3%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0.36</td>
<td>0.20</td>
<td>0.16 (0.12 – 0.20)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>0.33</td>
<td>0.42</td>
<td>-0.09 (-0.14 - -0.04)</td>
</tr>
<tr>
<td>Unsafe of incorrect</td>
<td>0.72</td>
<td>0.75</td>
<td>-0.03 (-0.08 – 0.03)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.19</td>
<td>0.24</td>
<td>-0.06 (-0.10 - -0.02)</td>
</tr>
</tbody>
</table>

p<0.007

It is possible that the failure to recognise that the correct response was missing, was mitigated for in the standard format by a degree of cueing. When the free-text box was selected to respond to the question in the non-standard format, compared with the options selected in the standard format, the student had a higher proportion correct, had a lower proportion unsafe, had a lower proportion incorrect of being unsafe, and had a lower proportion of ‘don’t know’ responses (Table 8.35), and this would be consistent with appropriate use of self-generated responses.
Table 8.35: Standard vs. non-standard format, free-text box selected

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Non-standard, text box selected</th>
<th>Difference (99.3%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0.36</td>
<td>0.70</td>
<td>-0.34 (-0.44 - -0.23)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>0.33</td>
<td>0.20</td>
<td>0.12 (0.03 – 0.22)</td>
</tr>
<tr>
<td>Unsafe of incorrect</td>
<td>0.72</td>
<td>0.54</td>
<td>0.18 (0.00 – 0.37)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.19</td>
<td>0.00</td>
<td>0.19 (0.17 – 0.20)</td>
</tr>
</tbody>
</table>

p<0.007

The self-generated responses using the free-text box when high certainty had a proportion correct was 0.96 with no unsafe responses. When the correct option was removed from the non-standard format questions the students had a lower proportion correct and a higher proportion unsafe (Table 8.36).

Table 8.36: Non-standard format, correct options present vs. non-standard format, correct options removed

<table>
<thead>
<tr>
<th></th>
<th>Non-standard, correct options present</th>
<th>Non-standard, correct options removed</th>
<th>Difference (99.3%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0.36</td>
<td>0.18</td>
<td>0.17 (0.10 – 0.25)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>0.34</td>
<td>0.43</td>
<td>-0.09 (-0.17 - -0.01)</td>
</tr>
<tr>
<td>Unsafe of incorrect</td>
<td>0.73</td>
<td>0.72</td>
<td>0.01 (-0.10 – 0.11)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.20</td>
<td>0.22</td>
<td>-0.02 (-0.09 – 0.04)</td>
</tr>
</tbody>
</table>

p<0.007

For non-standard format questions, there was a higher proportion correct and a lower proportion unsafe, and a lower proportion unsafe of those incorrect, and a lower proportion of ‘don’t know’ responses when the free-text box was used compared to when an option was selected (Table 8.37).
Table 8.37: Non-standard format, option selected vs. non-standard format, free-text box selected

<table>
<thead>
<tr>
<th></th>
<th>Non-standard, option selected</th>
<th>Non-standard, free-text box selected</th>
<th>Difference (99.3% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0.20</td>
<td>0.70</td>
<td>-0.50 (-0.61 - -0.39)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>0.42</td>
<td>0.20</td>
<td>0.21 (0.11 – 0.31)</td>
</tr>
<tr>
<td>Unsafe of incorrect</td>
<td>0.75</td>
<td>0.54</td>
<td>0.21 (0.02 – 0.40)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>0.24</td>
<td>0.00</td>
<td>0.24 (0.21 – 0.28)</td>
</tr>
</tbody>
</table>

p<0.007

As noted above in relation to the high certainty responses, it is not the higher certainty self-generated free-text responses that would create clinical risk occurring in practice, it is failure to recognise that the correct response is missing.

8.5.8 Change in responses between Test 1 and Test 2

A total of 19 students sat both Test 1 and Test 2. In Test 2, there were 35 questions sat by all candidates. The remaining questions were randomly selected. Due to a selection error, four candidates had some duplicated questions, and these responses were excluded from the analysis.

Analysis of responses of the 19 students to the 35 questions for internal consistency, gave a Cronbach’s α for correctness of 0.36, for certainty of 0.65, and for unsafeness of 0.32. This is a limited number of students and questions.

The cohort total responses for Test 1 (Table 8.38) and Test 2 (Table 8.39) are shown.
Table 8.38: Test 1 responses of the 19 students who sat both tests

<table>
<thead>
<tr>
<th></th>
<th>No certainty</th>
<th>Low certainty</th>
<th>Mod. certainty</th>
<th>High certainty</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>–</td>
<td>177</td>
<td>291</td>
<td>146</td>
<td>614</td>
</tr>
<tr>
<td>Incorrect not unsafe</td>
<td>–</td>
<td>103</td>
<td>75</td>
<td>15</td>
<td>193</td>
</tr>
<tr>
<td>Incorrect low unsafe</td>
<td>–</td>
<td>144</td>
<td>68</td>
<td>14</td>
<td>226</td>
</tr>
<tr>
<td>Incorrect mod. unsafe</td>
<td>–</td>
<td>98</td>
<td>75</td>
<td>9</td>
<td>182</td>
</tr>
<tr>
<td>Incorrect high unsafe</td>
<td>–</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>Don’t know</td>
<td>188</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>188</td>
</tr>
<tr>
<td>Total</td>
<td>188</td>
<td>540</td>
<td>513</td>
<td>185</td>
<td>1426</td>
</tr>
</tbody>
</table>

Table 8.39: Test 2 responses of the 19 students who sat both tests

<table>
<thead>
<tr>
<th></th>
<th>No certainty</th>
<th>Low certainty</th>
<th>Mod. certainty</th>
<th>High certainty</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>–</td>
<td>155</td>
<td>282</td>
<td>270</td>
<td>707</td>
</tr>
<tr>
<td>Incorrect not unsafe</td>
<td>–</td>
<td>70</td>
<td>61</td>
<td>30</td>
<td>161</td>
</tr>
<tr>
<td>Incorrect low unsafe</td>
<td>–</td>
<td>107</td>
<td>74</td>
<td>28</td>
<td>209</td>
</tr>
<tr>
<td>Incorrect mod unsafe</td>
<td>–</td>
<td>108</td>
<td>99</td>
<td>42</td>
<td>249</td>
</tr>
<tr>
<td>Incorrect high unsafe</td>
<td>–</td>
<td>16</td>
<td>13</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>Don’t know</td>
<td>50</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>456</td>
<td>529</td>
<td>372</td>
<td>1407</td>
</tr>
</tbody>
</table>
In Test 2 the students responded with a greater proportion of high certainty and a lesser proportion of low certainty and ‘don’t know’ responses. There was a greater proportion who answered correctly (Table 8.40).

Table 8.40: Change in responses with time, the difference Test 2 to Test 1

<table>
<thead>
<tr>
<th>Response</th>
<th>Proportion of which responses</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Difference, Test 2-Test 1 (99.8%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>All responses</td>
<td>0.43</td>
<td>0.50</td>
<td>0.07 (0.01 – 0.13)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>All responses</td>
<td>0.30</td>
<td>0.35</td>
<td>0.05 (-0.01 – 0.10)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>All responses</td>
<td>0.13</td>
<td>0.04</td>
<td>-0.10 (-0.13 - -0.06)</td>
</tr>
<tr>
<td>Low certainty</td>
<td>All responses</td>
<td>0.38</td>
<td>0.32</td>
<td>-0.05 (-0.11 – 0.00)</td>
</tr>
<tr>
<td>Mod. certainty</td>
<td>All responses</td>
<td>0.36</td>
<td>0.38</td>
<td>0.02 (-0.04 – 0.07)</td>
</tr>
<tr>
<td>High certainty</td>
<td>All responses</td>
<td>0.13</td>
<td>0.26</td>
<td>0.13 (0.09 – 0.18)</td>
</tr>
<tr>
<td>Unsafe</td>
<td>Incorrect responses</td>
<td>0.69</td>
<td>0.75</td>
<td>0.06 (-0.02 – 0.14)</td>
</tr>
</tbody>
</table>

*p<0.002

There was no difference in the proportion correct, unsafe, or unsafe or incorrect for any level of certainty (p>0.002 for all). Correctness and certainty appeared to increase in unison. There was a significant difference between tests in the proportion correct with OR 1.47 (95%CI, 1.22 - 1.77) (F=16.13, df=1,2380, p<0.0001). There was a significant difference between tests in the proportion ‘don’t know’ with OR 0.25 (95%CI, 0.18 - 0.36) (F=60.33, df=1,2380, p<0.0001). There was a significant difference between tests in certainty of responses excluding ‘don’t know’ OR 1.95 (95%CI, 1.64 - 2.31) (F=58.03 df=1,2151 p<0.0001) (Table 8.41).

Table 8.41: Proportion of certainty of responses, excluding ’don’t know’

<table>
<thead>
<tr>
<th>Test</th>
<th>Low certainty</th>
<th>Mod. certainty</th>
<th>High certainty</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.5% (541)</td>
<td>41.4% (516)</td>
<td>15.1% (188)</td>
<td>1245</td>
</tr>
<tr>
<td>2</td>
<td>33.7% (455)</td>
<td>39.0% (526)</td>
<td>27.3% (369)</td>
<td>1350</td>
</tr>
</tbody>
</table>
There was also no significant difference between tests in Unsafe (excluding 'don’t know’) OR 0.85 (95%CI, 0.70 - 1.04) (F=2.51, df=1,2150, p=0.11) (Table 8.42).

**Table 8.42: Proportion of unsafeness of responses, excluding 'don’t know'**

<table>
<thead>
<tr>
<th>Test</th>
<th>Not unsafe</th>
<th>Low unsafeness</th>
<th>Moderate unsafeness</th>
<th>High unsafeness</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65.4% (814)</td>
<td>18.2% (226)</td>
<td>14.6% (182)</td>
<td>1.8% (23)</td>
<td>1245</td>
</tr>
<tr>
<td>2</td>
<td>63.8% (861)</td>
<td>15.5% (209)</td>
<td>18.4% (249)</td>
<td>2.3% (31)</td>
<td>1350</td>
</tr>
</tbody>
</table>

Similarly, there was no significant difference between tests in any unsafe (excluding 'don’t know’) with an OR 0.85 (95%CI, 0.69 - 1.04) (F=2.40, df=1,2152, p=0.12) (Table 8.43), and no significant difference between tests any unsafe among incorrect (excluding 'don’t know’) with an OR 1.12 (95%CI, 0.77 - 1.63) (F=0.34 df=1,929 p=0.56) (Table 8.44).

**Table 8.43: Proportion of any unsafeness of responses, excluding 'don’t know'**

<table>
<thead>
<tr>
<th>Test</th>
<th>Not unsafe</th>
<th>Unsafe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65.4% (814)</td>
<td>34.6% (431)</td>
<td>1245</td>
</tr>
<tr>
<td>2</td>
<td>63.8% (861)</td>
<td>36.2% (489)</td>
<td>1350</td>
</tr>
</tbody>
</table>

**Table 8.44: Proportion of any unsafeness among incorrect responses, excluding 'don’t know'**

<table>
<thead>
<tr>
<th>Test</th>
<th>Not unsafe</th>
<th>Unsafe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.2% (195)</td>
<td>68.8% (431)</td>
<td>626</td>
</tr>
<tr>
<td>2</td>
<td>24.5% (159)</td>
<td>75.5% (489)</td>
<td>648</td>
</tr>
</tbody>
</table>

Where the relationship of levels of certainty to a correct response was concerned, there was no significant difference between the tests (F=1.21, df=2,2589, p=0.30). On the other hand, there were significant differences between levels of certainty for both Test 1 (F=61.84, df=2,2589, p<0.0001) and Test 2 (F=58.23 df=2,2589 p<0.0001) (Table 8.45).
Table 8.45: Odds ratio correct for levels of certainty

<table>
<thead>
<tr>
<th>Test</th>
<th>Certainty</th>
<th>Percentage correct</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low</td>
<td>32.9% (178/541)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>56.8% (293/516)</td>
<td>2.68 (2.09 – 3.44)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>78.7% (148/188)</td>
<td>7.55 (5.09 – 11.18)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>2</td>
<td>Low</td>
<td>33.8% (154/455)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>53.2% (280/526)</td>
<td>2.22 (1.72 – 2.88)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>72.6% (268/369)</td>
<td>5.19 (3.84 – 7.00)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

There was no significant difference between the tests in the relationship of levels of certainty to a unsafe response (F=1.24, df=2,2589, p=0.29). However, there were significant differences between levels of certainty for both Test 1 (F=41.94, df=2,2589, p<0.0001) and Test 2 (F=40.98, df=2,2589, p<0.0001) (Table 8.46).

Table 8.46: Odds ratio of any unsafe for levels of certainty for Test 1 vs. Test 2

<table>
<thead>
<tr>
<th>Test</th>
<th>Certainty</th>
<th>Percentage unsafe</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low</td>
<td>48.1% (260/541)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>28.5% (147/516)</td>
<td>0.43 (0.33 – 0.56)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>12.8% (24/188)</td>
<td>0.16 (0.10 – 0.25)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>2</td>
<td>Low</td>
<td>50.8% (231/455)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>35.4% (186/526)</td>
<td>0.53 (0.41 – 0.69)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>19.5% (72/369)</td>
<td>0.24 (0.17 – 0.32)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Continuing the pattern, further analysis revealed no significant difference between the exams in the relationship of levels of certainty to a unsafe of the incorrect responses (F=0.34, df=2,1268, p=0.71), and no significant differences between levels of certainty for either Test 1 (F=1.82, df=2,1268, p=0.16) and Test 2 (F=0.61, df=2,1268, p=0.54) (Table 8.47).
Table 8.47: Odds ratio of any unsafe within incorrect for levels of certainty for Test 1 vs. Test 2

<table>
<thead>
<tr>
<th>Test</th>
<th>Certainty</th>
<th>Percentage unsafe</th>
<th>OR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low</td>
<td>71.6% (260/363)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>65.9% (147/223)</td>
<td>0.77 (0.54 – 1.10)</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>60.0% (24/40)</td>
<td>0.59 (0.30 – 1.16)</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>Low</td>
<td>76.7% (231/301)</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>75.6% (186/246)</td>
<td>0.94 (0.63 – 1.40)</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>71.3% (72/101)</td>
<td>0.75 (0.45 – 1.25)</td>
<td>0.27</td>
</tr>
</tbody>
</table>

To summarise these Test 1 vs Test 2 comparisons there is an increase in correctness, with a reduction in levels of incorrect, unsafe and don’t know responses, and an increase in levels of certainty.

8.5.9 Student interview results

When using a general inductive approach, the analysis is overtly influenced by the researcher’s perspective, so I include a brief self-reflection:

My role, as a clinician, a senior lecturer and member of MBChB degree course policy committees, will have influenced my perspective on assessments and potentially students’ responses.

I had not at the time of the interviews, or to date, sat an assessment where certainty-in or safeness-of responses were specifically used. The only exception to this is testing the computer delivery for these studies, which I was doing with the intent of testing the software, for example deliberately getting some questions incorrect and unsafe.

This research programme developed from my experiences of advising students to guess if unsure in a test, but advising trainees to consider their certainty and consequences. Therefore, following a review of the literature and developing my own perspectives on this literature, I am now in a position of believing that the assessments need to be developed to better align to...
practice, and that certainty-in and safeness-of assessment responses reflects self-monitoring, and will go part way to meet this purpose.

At the time of the interviews, Study 1 was complete and Study 2 was part way through and had undergone an interim analysis. Provisional analysis of Test 1 and Test 2 of Study 3 had been undertaken to allow for sampling. I was aware of all these results, and had been considering the implications.

### 8.5.10 Interviews results

Fifteen students were invited to an interview: seven replied, six requested face-to-face interviews and one a videoconference. Interestingly, none of the students with highly certain highly unsafe responses from either test agreed to an interview.

The interviews were audio-recorded, and transcribed verbatim. Three electronic recording files had corrupted: two could not be used at all, but one had some retrievable data. Interviewer notes were available for these recordings, but did not contain verbatim quotes.

As the researcher after all the interviews were complete, I read the transcripts, while listening to the recordings, then re-read the transcripts. Then went through the transcripts and made notes on points raised. At the same time the interview notes for the interviews with no transcript were used to generate notes on points raised. Then commenced the development of themes from the notes of points raised. A re-read confirmed these themes and their relationship to the points raised. As this analysis progressed connections were made with literature and prior results.

### 8.5.11 Themes

The themes identified from participant interviews were:

1. Certainty-in and safeness-of response format MCQ tests were perceived as authentic to clinical practice.
2. Engagement was reduced by the context of tests as a formative only, research project additional to the standard education programme;
3. Responses selected included consideration of certainty-in and safeness-of responses.
4. Feedback on correctness for certainty and safeness for certainty was important. Certainty-in and safeness-of responses provides additional feedback compared to standard assessment that is useful to guide future learning.

5. Both the assessment process and feedback stimulated a personal review of self-monitoring.

6. There is a use for certainty-in and safeness-of responses in assessments to inform pass-fail progress decisions.

Each of these themes will now be described in more detail.

8.5.11.1 Theme 1: Participant’s perceptions of authenticity

The first theme to be described relates to perceptions of authenticity. As has been described in the literature review, self-monitoring, decision certainty and decision safety are all authentic within clinical practice. However, this does not mean that replicating similar aspects within the simulated environment of an MCQ test would be perceived as valid.

The link to subsequent practice was apparent.

I thought it probably reflected how I do act in a clinical setting. (Participant P)

You can draw parallels with [pause] if you are only moderately or have low certainty about one situation and you would go and get help. (Participant C)

I think it is probably more true to real life. (Participant C)

…in real life guessing doesn’t go very well. (Participant R)

The question format provoked students into the realisation that decision making in practice would come with associated certainty, the need to check with a resource or person, and consequences.

I guess that I hadn’t really understood that I will in two years’ time be needing to make decisions on my own. (Participant P)

…probably it’s not always appropriate to consult somebody else. (Participant P)
...TI or your house surgeon and they would discuss always come back and refer it somebody else and discuss it, and maybe the registrar would say, well, have you thought about this? (Participant P)

really important at the end of 5th year heading into junior doc (Participant C)

Despite the concept of certainty and safety consequences being authentic, this was still a paper-based simulated exercise.

...this isn’t a real person. (Participant P)

Finally, this student was looking further into the future, realising that they may be the more experienced one being asked for advice.

...probably one day I will be the person being consulted. (Participant P)

8.5.11.2 Theme 2: Participants engagement

As this was an opt-in research examination held in the evening, the students did not approach it as they would higher stakes assessments.

I was perhaps not invested emotionally, because it was not going to be linked to my score. (Participant P)

...before test in a medical school situation I would revise, whereas this was 5 to 7 pm and I had been in class for 12 hours. (Participant P)

...little less invested in getting all the details. (Participant P)

...if it is formative, people don’t really care. (Participant P)

... affected by being out of school hours, rushing through it. (Participant C)

I did not have any stake in the experiment. (Participant R)

Some did not approach this study with the purpose of answering as they would in authentic practice or a higher stakes medical school assessment.

...results back, and so that changed how I played the game. (Participant P)

I didn’t like to use it. (Participant C) (referring to the Don’t Know option)
If students were actually studying for the test at the time they were sitting, they would probably do better. (Participant C)

I was not going to get it right. I guessed. (Participant R)

I would have put fewer answers as being high certainty. (Participant R)

...generally I was uncertain....moderately low, I think I put high in there once, just to mix it up. (Participant P)

8.5.11.3 Theme 3: Participants considered how they selected their responses

Participants initially selected what they thought were correct responses. This may have involved looking for the correct response among the options, or looking through the options for cueing to the correct response.

...look at the stem – not sure, look at the options: can I relate any of the options to the stem? (Participant H)

...working out the stem and going through the answers. (Participant T)

I felt at some point I should choose the answer I thought it was. (Participant P)

Participants attempted to avoid unsafe responses such as errors of omission, not wanting to miss serious, high consequence diagnoses.

I would sway my answers to something more serious. (Participant P)

I would not want to come across as unsafe. (Participant C)

I am concerned about safety. (Participant T)

The level of certainty included considerations specific to the questions, but also specific to individual factors. For the students who were interviewed, question-related factors were considered first.

The certainty scale reflected the students' considerations of certainty.

I think it does reflect my certainty. (Participant P)

The process of considering certainty, may impact on certainty.
I think I maybe breed a bit of doubt in my head. (Participant H)

The level of certainty might depend on the ease with which the response is decided.

...process of elimination. Just go through the each of the options and come up with an answer that is glaringly obvious... go for high certainty. (Participant R)

I knew something I would usually put it under high certainty. (Participant R)

...the ones I did not know, or I was unsure of, I would put below, and I mean below more than moderate. (Participant R)

...you can work out the answers. I think it just added an element so my impression is that widens the confidence interval, you are kind of guessing, and things that you do get is more likely for lower certainty. (Participant R)

The perceived consequence (unsafeness) would also impact the certainty.

...lesser consequence if I got it wrong, I might gamble a little more and I was pretty sure. (Participant C)

One participant felt that the number of options did not affect their certainty, unless there were too many options, then it might.

...it certainty is different in the way you do it. Ignore the options, read the stem, what do you think is happening? (Participant H)

if there’s too many options that are potentially likely that would likely .... putting a high certainty answer. (Participant H)

Some topics were considered as being more likely to have highly unsafe responses, and this would have an impact on certainty.

...high risk scenario. I might be less likely to put high certainty. (Participant C)

...quite serious or life threatening, then I would be less certain about my answer. (Participant C)

Some topics were considered to require a high certainty response, as a clinician should act if authentic.
emergency question or more benign sounding question. It sort of rings alarm bells around, I should be highly certain about certain things. (Participant H)

The participants considered that their personal certainty might have a baseline or a limited range.

...it is sort of default put me into a low certainty answer. (Participant H)

...some people were all high or medium (Participant H)

...probably depending on my personality on the day (Participant R)

8.5.11.4 Theme 4: Participants considered the utility of the feedback

Students looked at different aspects, including correctness for certain, safeness, and the bigger picture.

I looked at the whole combination. (Participant R)

Definitely safety was more concerning. (Participant P)

...the certainty scale was the second thing I looked at. (Participant T)

Correctness of certainty, or rather incorrectness with high certainty, was a consideration.

No need to think why I came to the right answer, probably more important to think why I got the wrong answer. (Participant P)

I was particularly concerned about those if I had put the question as high certainty. (Participant C)

Also, that it was incorrect with lower certainty was less important.

I was uncertain so I didn’t really worry about how certain (Participant P)

A lack of unsafe responses at higher certainty was seen as reassuring.

I was happy to note that I didn’t get any thought I was right, definitely not right, this is dangerous. (Participant P)
…the only ones where you had higher unsafety was when there was lower certainty. (Participant H)

I went through and checked, say, any of the ones that I put above low certainty that got above low unsafe. (Participant H)

I think that by having the certainty thing, it kind of gives you reassurance that at least you know where you stand, and that is important in terms of safety. (Participant R)

I put down 15 of them as being high certainty, and none of them are unsafe so that was reassuring. (Participant R)

And higher certainty higher, unsafe responses was seen as concerning.

…pretty unsafe, pretty certain, and I thought oh no. (Participant P)
…knowing your safety levels was important. (Participant P)

…I scrolled through to find the high risk wrong answers. (Participant C)

Feedback on unsafe responses provides additional information, over and above standard assessment feedback, being useful to focus future learning

…it’s not just a wrong answer: it’s a wrong answer with repercussions. (Participant P)

You don’t know how wrong the wrong answer was. (Participant H)

…unsafe, and you think you are right, obviously you don’t get that level of data when you are wrong in a question. (Participant P)
…highly unsafe would definitely be something I would be concerned about: if that does reflect how I think, I might need to change that. (Participant P)

I think it is a little bit more informative. (Participant C)
…more informative, and it is probably better for learning, but there is probably more at stake with your answers and interpretations (Participant C)

According to the participants, the feedback associated with this format of assessment might help direct and motivate future learning.
...it really made me think. (Participant P)
I would quite probably quite often check that I would consult a reference (Participant P)
I think it was good at highlighting things that are going to be bread and butter and I don’t understand. (Participant P)
I know what I need to work on. (Participant P)
I did actually go look up sorts of lumps in the neck lymph nodes. (Participant H)
...those topics you need to particularly focus on. (Participant C)
...the holes in your knowledge are more obvious. (Participant C)
...going back through the topics that I didn’t do as well on. (Participant C)
...questions that I got wrong and work out why I may have selected that.
...verifies the things you don’t know. (Participant T)

8.5.11.5 Theme 5: Participants considered self-monitoring
The participants reflected on the rationale and rigour of their own meta-judgements. The process of certainty-in and safeness-of will introduce the concept of self-monitoring.

...you could think about why that was a terrible guess. (Participant P)

...make a decision, but you have also got to back it up, and I think it’s useful for learning, because it make you think, why do I think that I am positive? (Participant C)

I think it does give you insight. (Participant R)

8.5.11.6 Theme 6: Participants considered the future possibility of this format being used to inform progression decisions
In future safety, more than certainty, was considered to be an aspect that could be used in pass-fail decision-making.
...few people in our class who are naturally bold...they might be wrong but they don’t care, they are going to run with it, whereas I would always prefer to know the right answer. (Participant P)

I think that’s pretty reasonable... If it was going to be a dangerous answer, then it should be a low confidence guess. (Participant H)

So when they are counting, tallying up the whole question or something, they do actually take into account very unsafe questions. (Participant T)

Including certainty in the consideration was seen as useful for some circumstances, but not always.

...confidence may be taken into account for a borderline result. (Participant H)

...two relatively equal students, using different criteria for assigning confidence and perhaps the more confident get the same amount right. Are they perceived as better? I suppose they are...I’m not sure. (Participant H)

I think that certainty does sound quite important, in that it would imply understanding. (Participant H)

...you can pass from guessing... I think by having certainty thing you can kind of stop that from happening. (Participant R)

The nature of higher stakes leading to more engagement was raised again.

...if it is summative, then people target their learning. I think it would be helpful. (Participant P)

The results of the analysis of responses, comparison with standard assessments, the option-list and free-text responses, Test 1 and Test 2 responses, and student interviews have generated several results related to exploration of certainty-in and safeness-of, and these are discussed in the following section.
8.6 Discussion

Initially the results will be summarised in the context of the hypotheses and research questions. The results will be discussed related to feasibility of certainty-in and safeness-of responses, correctness for levels of certainty, and safeness for levels of certainty. Then the results related to a lack of effect of feedback on proportions correct or unsafe for given levels of certainty, risk-taking in responses, correlations other assessments, gender, and further interpretations of themes from interviews will be discussed. Following this, findings not related to the research questions will be discussed. Lastly, the strengths and limitations of this study will be discussed.

As a summary of the results for the hypotheses:

- Undertaking certainty-in and safeness-of MCQ responses as a format for an MCQ test with a larger number of questions covering more diverse content being feasible was confirmed
- Students of all ability levels, characterised by criteria, would demonstrate increasing correctness with certainty was unable to be confirmed or refuted
- Among students with higher ability, unsafe responses would be more likely as levels of certainty increase was unable to be confirmed or refuted
- Some, but not all, results for certainty-in and safeness-of MCQ response questions tests would correlate with OMS assessments results was confirmed, although not with strong correlations
- Gender would impact on certainty-in and safeness-of MCQ responses was confirmed. An impact of student ethnicity could not be confirmed or refuted.
- Elapsed time and feedback would lead to an improvement in response correctness and unsafeness for certainty was refuted
- Student responses would demonstrate risk-taking patterns was confirmed (partially)
- Student perceptions on approach to, feedback from, and utility of certainty-in and safeness-of MCQ responses would be positive was confirmed
As a summary of the results related to the research questions:

1. Can certainty-in and safeness-of responses be included in MCQ examination with a larger number of questions?
   - Yes, possible for a question bank for 180, with examinations of 75 items.

2. What is the relationship between response certainty and correctness?
   - Correctness increased with certainty.

3. What is the relationship between response certainty and the potentially unsafe responses?
   - Unsafe responses decrease with increasing certainty.
   - Unsafe responses at high certainty are rare.

4. How do these relationships of certainty with correctness and unsafeness vary with ability (number correct) groupings?
   - Unable to analyse.

5. How do certainty-in and safeness-of responses compare with other assessment results?
   - Correlation with few faculty assessments especially related to format and content.

6. How do certainty-in and safeness-of responses and other assessments compare between gender and ethnicity groups?
   - Males had higher certainty; females had more unsafe responses with increasing certainty.
   - Unable to analyse for ethnicity groups.

7. How do certainty-in and safeness-of responses assessments vary in questions modified to detect risk-taking behaviour?
   - Increased number correct, decreased incorrect, unsafe and 'don’t know' responses, and increased level of certainty.
   - No difference in proportions correct, unsafe, or unsafe of incorrect for level of certainty.
• Removal of the correct option, and the need to generate the correct response, leads to less correct and more unsafe responses.
• Self-generated responses are more likely to be correct and are less likely to be unsafe, especially those with high certainty.

8. How does time elapsed/feedback alter response patterns for certainty-in and safeness-of responses?
  • Increased number correct, decreased incorrect, unsafe and ‘don’t know’ responses, and increased level of certainty.
  • No difference in proportions correct, unsafe, or unsafe of incorrect for level of certainty.

9. What do students perceive regarding their approach to, feedback from, and utility of, certainty-in and safeness-of response MCQs?
  • Certainty-in and safeness-of response MCQ format tests were perceived as authentic to clinical practice.
  • Engagement was reduced by the context of tests as a formative only, research project sat in addition to the standard education programme.
  • Responses selected included consideration of certainty-in and safeness-of responses.
  • Feedback on correctness for certainty, and safeness for certainty, was important.
  • Certainty-in and safeness-of responses provide additional feedback compared to standard assessment that is useful to guide future learning.
  • Both the assessment process and feedback stimulated personal review of self-monitoring.
  • There is a use for certainty-in and safeness-of responses in assessments used to inform pass-fail progress decisions.

8.6.1 Feasibility of certainty-in and safeness-of responses

There were no feasibility issues apparent, or raised by the participants, in completing the certainty scale.
The level of agreement in the safeness of responses between the five panellists, over 180 questions and 2345 incorrect responses, was Kappa 0.11. This represents a low level of agreement. In addition to being resource-intensive, this does create an issue. Given the lower level of agreement, either more dialogue among the panellists and/or more panellists would be required if high-stakes decisions were to be informed by the levels of unsafeness of responses.

In addition, and relating to feasibility, the attempt at standard-setting at multiple levels revealed differences in variance of Angoff judgements between years, and greatest variance for Year 3 judgements. Again, this implies that either more dialogue among the panellists and/or more panellists would be required for setting pass standards that would be consistent at multiple stages within a course.

8.6.2 Correctness for levels of certainty

As in Study 1 and Study 2, correctness increased with certainty, and this was demonstrated for the cohort as a whole. This finding was confirmed for subgroups defined gender. There were insufficient numbers of students above standard to use this as a metric for group characterisation. Likewise, this analysis was not possible for student ethnicity grouping.

8.6.3 Safeness for levels of certainty

Like Study 1 and Study 2, the proportion of unsafe responses decreased with increasing certainty. Unlike in Study 1 and Study 2, there was no change in unsafe responses amongst the incorrect responses with increasing certainty. As certainty increased, the odds of unsafe responses reduced, and unsafe as proportion of incorrect responses did not change. This pattern was seen in all subgroups analysed.

What was different about Study 3 compared to Study 1 and Study 2? The main difference was a greater number of questions and more diverse content. It is possible that Study 3, with a greater number of students, questions and content better reflects reality. However, this does need further study. In Study 3 there was a low number of highly unsafe, highly certainty responses: four, from four different students. Again, as for Study 1 and Study 2, this is reassuring. Further, the free-text box responses had fewer unsafe responses.
8.6.4 Lack of effect of feedback on proportions correct or unsafe for given levels of certainty

The first possibility to consider in why feedback had no detectable effect is that neither correctness nor safeness for levels of certainty is amenable to change. This is unlikely, as Study 1, Study 2, and Study 3 Test 1 demonstrated differing levels of correctness and safeness for given levels of certainty across students with different levels of ability and experience. Therefore, one could postulate that for an individual with changing ability and experience, their correctness and safeness for given levels of certainty would change.

Context-dependent meta-judgements can be learnt and improve with feedback. Feedback highlights the gaps between current levels and what is required, and so has the potential to enhance learning. Self-efficacy is responsive to change, meaning that self-regulation, certainty, self-assessment and self-efficacy may be enhanced by feedback. If a change was expected but not detected, why might this be? It is possible that no effect was detected because the assessment tool could not detect change in this number of students. The number of students was low, and given they were volunteers, it is possible that they were not a representative group.

Feedback might lead to change, but no effect was present given the timing of this study. Feedback may take more time to work, or require more episodes of feedback, or perhaps students would need additional experience of the applicability of the feedback. Calibrated accuracy meta-judgements remained stable over the course of a semester, raising the question: is prompting students to think about their performance too passive an approach to altering correctness or safeness for levels of certainty? It was postulated that to improve meta-judgement skills, more intensive, explicit guidance with practice is required. This additional practice experience may not have occurred to re-enforce the feedback.

The feedback might not have been of sufficient quality. Many students do get feedback, but the effect of feedback depends on its quantity, quality and use. Feedback should include information pertinent to: task; cognition factors; self-regulation; self-management; appropriate timing, which may be different for different factors; positive and negative factors; and individual and group. Topics, levels of correctness, associated levels of certainty, highlighting unsafe high certainty responses and number correct standards, and no normative data were
included in the feedback. These metrics are seen to be advantageous to include in this type of feedback and feedback is most powerful when it includes and highlights mis-interpretations. Therefore, the content of the feedback provided to students in this study seemed appropriate for guiding further learning.

Was the lack of change between Test 1 and Test 2 because of student-specific factors? If the students are not motivated and do not take feedback seriously, then the efficacy of feedback could be reduced. Students need to be motivated and accepting of feedback, as certainty in incorrect responses can be resistant to change. The logic behind this is that deficits that are identified should lead to remediation. Improvement in self-assessment is limited, even when feedback is provided. The degree to which people will alter perceived self-efficacy on the basis of feedback depends on several factors including difficulty, effort, aid, and situation. Even when gaps are recognised, there may not be sufficient motivation to learn. This, self-assessment is one of the many factors that may influence self-directed learning.

Could improvement and subsequent decay in knowledge and self-monitoring with appropriate certainty have occurred between Test 1 and Test 2? Decay may occur differently for knowledge compared to self-monitoring with appropriate certainty. This leads to students displaying inappropriate certainty in the face of reducing knowledge retention.

8.6.5 Risk-taking

The presence of the free-text box when the correct response was present made no difference to certainty or unsafeness of responses. When the free-text box was present and the correct response had been removed, only 58% high certainty responses were correct, and 21% were unsafe. Higher certainty unsafe responses would be consistent with risk-taking in clinical practice. However when the free-text box was used to self-generate the response with high certainty, these were nearly always correct and never unsafe. It appears that it was not the self-generation of the response which created the unsafeness; rather, it was students not recognising that the correct response was missing.

Self-generated responses could be like SAQ responses. Cueing and chance are purported to increase the probability of a correct response in an MCQ compared
with SAQs. Regarding certainty, though, there have been contradictory results. SAQ responding was found to be better than MCQ responding for 'awareness of performance' as assessed by the student moving on rather than reviewing their response. No differences in certainty rating or correctness were detected between SAQs and MCQs, although there were differences within each item type when comparing correct and incorrect responses. The effect size for item type was less than that observed for content. Contrary to this, SAQ-predicted scores are lower than actual scores for SAQs, but not MCQs. Examinees were less likely to defer answering SAQs compared with MCQs. The difference between first-time answers and second-time (deferred because of uncertainty) answers is greater for SAQs and MCQs. In comparing confidence to performance, SAQs have a higher correlation than MCQs. First-time response rates were higher for MCQs, implying that SAQs are responded to with certainty, but on a second round review no new information is available to prompt a response. Certainty in a self-generated hypothesis and knowledge is generally reduced, but the act of self-generating responses takes effort, which is associated with an increase in certainty, without an equivalent increase in correctness. There are complex interactions for the MCQ and SAQ formats when considering certainty-in and safeness-of responses, and this could be an area for further exploration.

Could the MCQ format, the MCQ format used in this study and/or the MCQ format the students were used to, promote risk-taking responses? If the format of the assessment were to promote risk-taking, this would be a concern. Faced with a MCQ, when the candidate immediately knows the response to look for in the options, they will also recognise when it is not present, and as in this study, make a free-text response, and those made with high certainty are nearly always correct and never unsafe. However, when a candidate does not recognise the absence of the correct response, they might search the options and be cued to an incorrect, and potentially unsafe, option. It is possible that repeated exposure to certainty-in and safeness-of response MCQs might alter this response and/or strategies. This study was a one-off and it is possible the prior MCQs without certainty-in or safeness-of responses being taken into account might influence the students’ responses and strategies.
MCQ response options should include either a 'don’t know' or a 'no certainty' option, to both reflect authentic practice and to avoid any potential subliminal promotion of risk-taking.

8.6.6 Correlations other assessments, gender and ethnicity

Correlation with other assessments seems to relate mostly to format. The Retained Knowledge Test at OMS is an MCQ test covering the entire curriculum, that is not sat under examination conditions. The correlations for proportion correct and inverse correlations with proportion unsafe, but not the outcomes in including certainty, would be explained by similarity in content, format and administration. Correlations being higher between formats implies that the outcomes may be related to different individual characteristics *, such as correctness, correctness for certainty and safeness for certainty.

In Test 1, males were more certain in their responses and made fewer unsafe responses than females. Among incorrect responses, there was no difference in unsafe responses between genders. Females generally perform better in medical school assessments *, but MCQ is one area where there is more variation in results. Males score higher in true-false-abstain MCQs *, but not in all MCQs *. An MCQ format that allows for partial knowledge (subset selection), broadly aligned to certainty, revealed no gender difference *. Gender differences has been variable for response and scoring systems using certainty. Differences due to gender were reported as greater for conventional scoring than for certainty-based assessment (CBA) scoring *; no gender difference on formative-only or summative exams *; and gender differences present for CBA *. Males are more certain and overconfident on CBA scoring *. Why are there differences across formats, and the results shown in this study? That males had higher certainty has been reported by others, but the unsafe responses is a new finding. When females did not select the correct response, they were more likely to select an unsafe response, and in doing so, did so with higher certainty. This is a further area requiring confirmation and exploration, including the reliability of the safeness judgments made by panellists.

Analysis of responses by student ethnicity was not possible in a meaningful way in this study. Ethnicity differences in medical school exams have been reported *, with non-whites doing less well in UK Medical Schools. This is an area that would benefit from further exploration.
8.6.7 Further interpretations of themes from interviews

The development of certainty-in and safeness-of response MCQs had been developed to reflect self-monitoring of clinical decisions in practice. Participants perceived these questions were authentic to clinical practice. One would have expected this increase in authenticity to improve motivation. However, engagement was reduced by the context of tests as a formative-only, research project sat in addition to the standard education programme. The influence the stakes of the assessment had on motivation may have outweighed the improved authenticity. The stakes of the assessment will also affect the responding. Given the lack of stakes, candidates may have answered questions with higher certainty than they would normally have.

The participants’ opinions on how they selected responses included consideration of certainty-in and safeness-of responses. The results from Study 1, 2, and 3 quantitative analysis, and the background literature, led to a model for how the selections for certainty were made (Figure 8.1). The student will have a baseline level of certainty for this type of task. The student may consider their knowledge and/or experience for this topic, real and simulated, actual and perceived, at this moment, for this assessment format, and the given the assessment stakes and this will modify their level of certainty accordingly. This may be influenced, dependent on the item format, by the means that the response was first generated such as cueing from the list or subset selection. Knowledge and/or experience of potential adverse consequences and/or the urgency of situation, real and simulated, actual and perceived may also be considered, and this will modify their level of certainty accordingly. This may include the consequences of actions and inactions. Hence, the baseline certainty level may be influenced by perceptions of capability and consequences, but not necessarily both or either.
Figure 8.1: Model to illustrate self-monitoring of item responses in assessment

Self-monitoring as reflection-in-action and assumption of authentic clinical context

- Medical student making a response decision
- Personal level of certainty in general
- Perception of capability related to topic and specific scenario
- Perception of consequences related to topic and specific scenario (urgency and safety)
- Medical student response and degree of certainty linked to assistance or resources

Figure 8.1: Model to illustrate self-monitoring of item responses in assessment
Koriat proposes a conceptualisation of certainty in an assessment task as having two stages, each with biases. The first stage involves searching one’s knowledge through to when an answer is chosen. During this stage a bias occurs involves favouring positive rather than negative evidence. The second stage involves the evidence being reviewed, and certainty in the chosen alternative is assessed. The bias in the certainty assessment is a tendency to disregard evidence inconsistent with the chosen answer.

Many terms such as knowledge, cognition, memory, comprehension and then the same terms with a meta-prefix, occur in the literature. The first group of terms relates to facts, information, and thought and reasoning skills acquired through experience or education. The meta-group of terms is a personal judgement on the first as self-awareness. Definitions of meta-judgements have been vague, including self-awareness of knowledge of the task and problem solving skills, and the self-awareness of all cognition.

The participants’ perceptions were that people have a certainty level for a topic, on this day, and on this assessment format, given the assessment stakes. Some people are generally more certain and others more cautious, raising the possibility of an intrinsic level of certainty, which varies between people for same question and level of knowledge. Factor analysis has led to a ‘confidence’ factor being postulated, with confidence being a separate trait lying between personality and cognitive ability. The possibility of certainty of the person (in general) and certainty for a specific question response has been proposed.

Ease of recall to generate a response affecting certainty has been widely reported, much based on the work of Tversky and Kahneman, including errors in clinical decision-making. The easier it is to generate a response, the higher the certainty for that response, even when it is incorrect.

The implications of knowledge and/or experience of potential adverse consequences, real and simulated, actual and perceived, does contribute to the certainty decision. This was discussed in section 6.5.3 as negative knowledge, and explains some of the findings related to higher certainty unsafe responses in Studies 1, 2 and 3.
The knowledge and/or experience of urgency of clinical practice, real and simulated, actual and perceived demonstrates awareness that some decisions need to be made efficiently. This links to section 3.7.2.2, highlighting the difference between being certain when correct, and being correct when certain. It may link to the already highlighted issues of the spectrum of more efficient performance and safer performance⁵, and demonstrates how certainty when correct (self-monitoring for efficiency) and correct when certain (self-monitoring for safety) are different. Both were considered by the participants in the current study.

Although the feedback on correctness for certainty and safeness for certainty was seen as important, responses in Test 2 did not demonstrate an increase in correctness for certainty or a reduction in unsafeness for certainty. Other research with greater statistical power and higher stakes have found an increase in correctness high certainty answers following feedback, however⁶.

The views of certainty-in and safeness-of responses provide additional feedback compared with standard assessments useful to guide future learning and inform plans on the usefulness of feedback (section 3.2.1.2). It is more likely that the feedback was perceived as truly useful, but that the benefits of the feedback were not detected by the analysis of Test 1 - Test 2 differences.

Feedback is a significant part of assessment⁷, and is one of the most effective strategies to guide future learning⁸. Feedback and prior certainty responses will allow students to focus on future learning activities⁹. Feedback in light of certainty can influence future responses¹⁰ and it has been recommended by others that for feedback for learning, certainty should be included in the assessment¹¹,¹². Certainty will allow for better direction of feedback, with most notice taken of high certainty incorrect responses¹³. Feedback that provides only the correct answer to an incorrect response may have no effect on learning¹⁴. Adding safeness re-enforces that it is not so much the incorrect, but the unsafe incorrect that is more important, especially for high level certainty. The effect of the timing of unsafe incorrect answer feedback along with certainty, which in this case was two to three weeks after the assessment, was not investigated in this research. Previously, when considering feedback to correct errors, delayed feedback may be optimal compared to that which may be instant¹⁵. The levels of safeness, certainty and the question topic were included in the feedback provided to students in this
study and they perceived it as being helpful; to direct further learning. Elaborative feedback provides further substantive information, in this case, about certainty and safeness. However, more feedback does not always lead to better follow-up test results. Certainty in incorrect responses can continue, and even increase after feedback, if the feedback is not remembered. Self-assessment, as response accuracy, has been reported both to improve and not improve following feedback. Conversely, people may correct their responses irrespective of feedback they receive. The students who most need to improve are not always those most willing to. Certainty will alter the effect of feedback and certainty is a useful index of learning behaviour. as it reflects more than just knowledge; it also reflects affective and motivational states. The effectiveness of feedback is influenced by the degree of certainty in a response. High certainty errors can be less likely to be corrected. Durability, the likelihood that a response will be repeated at a future time, and certainty, are positively related. This holds true for erroneous and correct responses.

Both the assessment process and feedback stimulated a personal review of self-monitoring in this study. Although feedback is a significant contributor to the assessment purpose of guiding learning, including that related to self-monitoring, guiding learning can start before, and continue during, an assessment, before feedback can occur. Self-monitoring itself may drive self-regulated learning. Feedback should relate to both the task and to the self-monitoring related to that task, guiding ongoing self-monitoring.

Discrepancy is the error in the meta-judgement. When knowledge and certainty are high and discrepancy low, feedback and further study time can be directed to other areas, thereby improving learning efficiency. An aim would be that feedback will lead to a shift from incorrect unsafe high certainty responses to incorrect not unsafe low certainty responses, and from correct low certainty responses to correct high certainty responses. Feedback should allow students to assess their own insightfulness, decreasing certainty, even if errors are not corrected. It is instructive for candidates to think about the occasions when they make mistakes with high certainty, why they failed to see the risks they were taking.
Interview participants generally reported that they considered correctness for levels of certainty and safeness for levels of certainty separately. The categorisation of typical misconceptions in knowledge can be useful. Identification of rules underlying misconceptions can help learning through feedback. Feedback should allow for the correction of errors, with the seriousness of errors being highlighted as important through feedback provided.

Errors can be stratified by seriousness, which may be straightforward for non-complex mathematics questions. Non-serious errors are improved by comprehensive feedback, but serious errors can be relatively unaffected by feedback.

There is a potential use for certainty-in and safeness-of responses in assessments in informing pass-fail progress decisions. As noted in section 2.6.1, the environmental scan, knowing one’s limits in safe clinical decision-making is part of clinical practice, and should be introduced to medical students. The participants saw a greater value for using certainty-in and safeness-of MCQ responses assessments to guide learning than to inform pass-fail progress decisions. If used to inform pass-fail progress decisions, it was seen as offering additional information, and would require considered implementation, as per Study 2.

Following the analysis and interpretation of the responses by the participants, in light of Studies 1-3 data and the prior literature, the usefulness for certainty-in and safeness-of responses, especially in the context of self-monitoring of clinical decision-making in assessment and student learning, is supported. Further, the process of considering the development of decision certainty allows for guiding learning as to how to make better decisions, not just at the initial response level, but for certainty decision as well.

8.6.8 Findings not related to the research questions

The same pattern of internal consistency being highest for certainty, then correctness and lowest for unsafeness was seen again. However, once again these analyses were limited by a low number of students sitting Test 2 and a low number questions common to all students. This is an interesting result and extrapolation as a stand-alone result must be limited. However, Studies 1-3 have revealed a similar pattern, and this would fit with the concept of a general and
context dependent factors affecting level of certainty. This will be reviewed in the discussion sections of the last study and also related to all studies in this thesis in section 10.3.2.

8.6.9 Strengths of this study

Test 1 was in the setting of a larger examination in terms of item numbers and content. This allowed further analysis, including comparisons to other assessments results and student demographic factors such as gender.

The interview analysis was new to Study 3. The use of a thematic analysis approach as considered for use in medical education does have some strengths. The thematic analysis method is flexible and not confined by research paradigm, and is applicable to a post-positivist approach. The planned analysis and that which was possible aligned to the six phases described, although some were completed more so than others, did identify patterns in response. The overall goals of the research, as opposed to the interviews questions themselves, did reflect the use of certainty-in and safeness-of MCQ responses in a framework of self-monitoring, to use assessment to guide learning and inform progression decisions.

The responses from the interviews, when aggregated to the quantitative analysis from Studies 1-3 and the background literature, did align sufficiently to help inform consideration of in the development of certainty decisions. The application of the researcher’s knowledge interest to the process is part of the deductive thematic method.

8.6.10 Limitations of this study

The processes related to the safety judgements were resource intensive. A significant amount of staff time was invested in discussing safety in practice and then undertaking the judgements.

As a result of this study, it was apparent that students may not approach the questions and their responses (option selection and/or certainty) as they would medical school higher stakes assessment or authentic practice. This may make generalisations based solely on this data difficult.
Students did not receive identical questions in Test 1. Therefore, the total correct and total unsafe responses may be affected by question pool faced. No correction for true item difficulty or potential for unsafeness was possible. Item response theory analysis may help with this, but the number of students sitting would not be sufficient to undertake such an analysis in a meaningful way.

The number of students undertaking Test 2, and also sub-group analysis for Test 1, was limited by the low response rate. For example, Year 2-5 above and below standard leads to 8 sub-groups, for less than 100 students in total.

There may be several causes for the limited number of interviewees and, within those, a limitation of response profiles, with no students with higher unsafe higher certainty responses participating. It is unlikely that the themes uncovered were not correct, but that themes related to making high certainty incorrect and unsafe responses were not revealed.

A factor which may have influenced agreement to be interviewed or interview responses was that I, as the interviewer, was a staff member who was involved in student assessment at an individual, module and OMS level. In this study, the students may have considered disclosing sensitive information, which may have included approaches to and opinions of current assessments and/or wider education within OMS. Procedures aimed to make the interviewee feel as comfortable as possible were undertaken, including a convenient location and time, and a 'private' room or network link. Despite reassurances to students in the invitations, the position of the researcher may have influenced the response rate and potentially, their responses. Although there is no requirement that absolute number or proportion of responses be required to include in a theme, a lack of responses does raise the possibility of anecdote rather than a pattern of responding informing the themes.

On top of the potential limitations in number of responses, the corruption of some recording files would have further added to the possibility that some perspective may not have been covered and been reflected in the themes.

Some information in the recordings or the data may not have been used. The recordings were transcribed by an administrative assistant, rather than by myself as a researcher. The majority of the data upon which the analysis was
based was that of the transcripts so any non-verbal or unheard utterances may be lost. Interpretations as a single researcher does place a significant limitation on the analysis and subsequent conclusions drawn from the qualitative analysis. Although the position of the researcher and engagement in the known literature, and in this case studies as part of the thesis, as a strength of the analysis, it can also be considered a weakness, in that this may be constrain and narrow the analysis and results. The process followed in this study did not produce a final overall thematic map.
Chapter 9. Study 4
9.1 Introduction

9.1.1 Building on Studies 1-3

Building on Studies 1-3 and prior research, which have found the potential utility of certainty-in and safeness-of responses in an MCQ format with findings including an increase in correctness for levels of certainty, and variable results regarding unsafeness, there are areas identified for further development which are the focus of Study 4. These are exploring resource implications of including increasing numbers of students and questions, working with a higher stakes assessment, including of students with a wider range of ability, and looking at the effects of removing the 'don't know' option (DKO).

9.1.2 Developments addressed in Study 4

The practicalities of convening a panel to evaluate the questions for standard-setting and safeness judgements have significant resource implications. For Study 4, a less resource-intensive means for standard-setting and safeness judgements was investigated. Next, the assessment described in Study 2 included results obtained under high-stakes conditions and Study 3 was larger scale in terms of questions, if not students. For Study 4, undertaking certainty-in and safeness-of responses in a large scale (questions and students) MCQ assessment with high-stakes was investigated.

Study 1 and Study 3 were open to several year groups of students. However, participation in these studies was voluntary, so the range of student experience levels and ability was limited, thereby risking response bias. In setting a progress test, it is usual for all students from all year groups to sit the same examination at the same time. It is expected that as students pass through the course, their ability, and therefore scores, will improve. Therefore, for any given progress test there are students with a diverse range of experiences and abilities sitting an identical test. For Study 4, undertaking certainty-in and safeness-of responses in such a progress test context was investigated.

Finally, Studies 1-3 all included a DKO. There are advantages and disadvantages to including the DKO. Removal of the DKO can be justified, as it removes the variability of individual likelihood to use the DKO, which is not solely dictated by ability. To candidates, there is advantage to answer every question, even if
completely ignorant ». Even so, some cautious candidates will still leave some answers blank, potentially being unfairly penalised compared to peers ». This is because a requirement to answer all questions in the face of ignorance should not be part of doctors’ practice or education ». When selecting a DKO, respondents underestimate the likelihood of being able to answer correctly, either due to the scoring system and/or their personality ». If forced to answer a previously selected DKO, examinees are more likely than chance to be correct ». For Study 4, the DKO was removed and functionally replaced with a ‘no certainty’ descriptor for any given response.

9.1.3 Hypotheses and research questions

The hypotheses for Study 4 became:

- Less resource-intensive means for standard-setting and safeness judgements would be feasible.
- Undertaking a large scale, high-stakes progress test using certainty-in and safeness-of responses would be feasible.
- Students of all ability levels, characterised by specified criteria, and experience levels, would demonstrate increasing correctness with certainty.
- Unsafe responses, as a proportion of incorrect responses, would be more likely with increasing certainty for increasing levels of ability.
- Students’ ‘no certainty’ responses would be more likely to be correct than chance.

The research questions for Study 4 became:

1. Are there alternative means for standard-setting and safeness judgements?
2. Can certainty-in and safeness-of responses be included in a large scale high-stakes progress test?
3. What is the relationship between response certainty and correctness?
4. How do these measures of certainty and correctness vary with year group and ability?
5. What is the relationship between response certainty and safeness?
6. How do these measures of certainty and safeness vary with year group and ability?

7. What is the likelihood of correctness and safeness for responses marked as 'no certainty'?
9.2 Methods

9.2.1 Design

The data for Study 4 came from the regular administration of a retained knowledge (progress) test (RKT) delivered in 2015 as described in section 4.4.6.1, but briefly, each test administration involved 150 MCQs taken from a pool, sat by all medical students in Years 2-5. In addition to giving a selected question response, the student indicated their level of certainty in their answer as no, low, moderate, or high, with descriptors associated with each of these.

9.2.2 Question pool

9.2.2.1 Question pool content and question review

The questions for the RKT were taken from the OMS MBChB Assessment Item bank. Each MCQ incorporated a range of five to sixteen options. The content related to any aspects that occur as part of the core curriculum anywhere through to the completion of the course.

Standard question review was undertaken; that is not part of this thesis. Each question had been reviewed by one or more clinicians within the three years prior to question use. Three hundred questions were selected for the two RKTs of 150 MCQs each in 2015. No further content review was undertaken.

9.2.2.2 Standard setting process

Standard-setting methods other than Angoff were considered to replace the Modified Angoff method that had been used in Studies 1-3. Other methods for setting and equating standards in written examinations were considered. Many of these methods are resource-intensive. The resources used usually include time by a panel of judges and/or statistical expertise and software for analysis.

The search for a less resource-intensive, but acceptable standard-setting method for an assessment that was not a single-attempt barrier (very high stakes) led to Taylor’s modification of the Cohen method. This was chosen as it produces results equivalent to other methods, and has significantly lower resource
requirements. Minimum standards for each year group were set following the scoring based on total number correct.

9.2.2.3 Question classification for safety process

In the preceding studies, a panel of clinicians had been used to make judgements on the level of safeness of incorrect responses. In looking for other methods that would be less resource-intensive, none were identified: judgements by a panel of clinicians being the only method used by others. Data from authentic practice are not available on the actual outcomes of answers given that are incorrect and/or potentially unsafe. Realistically, there is no alternative to relying on the views of a panel of clinicians in order to make safeness categorisations for the responses.

Results from Studies 2 and 3 had demonstrated relatively low levels of agreement among the panellists with regard to the safety judgements. Therefore, a large number of panellists would be required.

Safety categorisation could not be done for these tests. Assembling a panel of sufficient size and experience with the time to review and categorise approximately 2000 distractors was not feasible.

9.2.3 Question selection for use in the tests

There were 150 MCQs chosen for each test. These were balanced for OMS MBChB degree course Curriculum Map categories as part of the blueprint for the test. No further categorisation or blueprinting was done as part of this study.

9.2.4 Subjects: candidates, recruitment, consequences/stakes

The RKT is a standard part of the normal curriculum for the Year 2-5 students. The purpose and consequences of the RKT are described in section 4.4.6.1. Briefly, all Year 2-Year 5 students were expected to complete the RKT as part of the standard assessment programme. Failure to do so, or have delayed or no completion approved was seen as a possible marker of failing to engage in educational activities and, as such, was notified to the relevant Student Progress Committee. There was no consequence of the number correct or certainty selections made, unless the number correct was less than chance or all the
certainty selections were the same. Again this was considered as a possible marker of failing to engage in educational activities.

9.2.5 Delivery format

Questions were delivered using the Learning Management System (LMS), MedMoodle. The front page of the test included a reminder to the students of the descriptors for levels of certainty (Table 9.1). These descriptors were developed from those used in Studies 1-3.

The administration of the test is as described in section 4.4.6.1. Briefly, the RKT was available online within a two week window, at a time and location convenient to themselves.
Table 9.1: Certainty descriptors

<table>
<thead>
<tr>
<th></th>
<th>No certainty</th>
<th>Low certainty</th>
<th>Moderate certainty</th>
<th>High certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I have no or insufficient experience and/or knowledge upon which to base a response.</strong></td>
<td><strong>My answer is:</strong> effectively a guess.</td>
<td><strong>My answer is:</strong> based on limited information.</td>
<td><strong>My answer is:</strong> based on partial information.</td>
<td><strong>My answer is:</strong> based on sufficient information.</td>
</tr>
<tr>
<td><strong>I would need to consult a colleague, clinician, or references prior to considering any response.</strong></td>
<td><strong>I would need to consult a colleague, clinician or references for assistance in formulating my response.</strong></td>
<td><strong>I would need to consult a colleague, clinician or references to confirm the appropriateness of my response.</strong></td>
<td><strong>I would have no need to consult a colleague, clinician or reference in order to make a response.</strong></td>
<td></td>
</tr>
<tr>
<td><strong>In an authentic healthcare situation, I would require education to respond.</strong></td>
<td><strong>In an authentic healthcare situation, I would require direction to respond.</strong></td>
<td><strong>In an authentic healthcare situation, I would require confirmation to respond.</strong></td>
<td><strong>In an authentic healthcare situation, I would be able to respond.</strong></td>
<td><strong>While I may consult a colleague or clinician, this is because they are required to undertake further action, not to educate, direct or confirm the candidate’s response.</strong></td>
</tr>
</tbody>
</table>

303
9.2.6 Information for students

The instructions to students were circulated at the start of the year, and on the front page of the test:

*The Retained Knowledge Test (RKT)*

The main purpose of the RKT (formerly known as the Progress Test) is to allow you to monitor and reflect on the knowledge that you accumulate as you progress through the MBChB course. The Medical School expects you to engage meaningfully with this process: your Student Progress Committee will be informed if you do not.

RKTs are simultaneously administered to all Y2 – Y5 students twice a year throughout the course. Each RKT comprises multiple-choice questions (MCQs) covering content appropriate to all levels of the course. The tests are delivered over a two-week period via MedMoodle to allow you to fit this in around your other commitments. Try and answer all 150 questions in one sitting: this will give the best snapshot of your current knowledge and insightfulness, and is excellent practice for sitting MCQs under exam conditions.

Each question requires two responses: selecting the correct answer, and rating how certain you are that you have selected the correct answer. Resist the temptation to look up answers, or to later change your degree of certainty. The intent is for you to develop the insightfulness required for good clinical practice, and to be able to recognise and accept situations where you have only partial knowledge of a problem or its solution. For descriptors and information related to certainty levels, see the table.

The certainty descriptor table (Table 9.1) for students was provided.

9.2.7 Feedback to students

The minimum standards for total correct for each year group were given to students as part of their feedback. Individual feedback included total correct, proportion correct for each level of certainty and proportion correct by curriculum discipline/domain categorisation. Additional feedback given to all students included the proportion correct for each level of certainty by year group, which could be considered normative.  

304
9.3 Analysis

9.3.1 Response and scoring reliability

Score reliability is an important aspect of testing, and for written tests the internal consistency is estimated by use of Cronbach’s $\alpha$.  

9.3.2 Analysis linked to research questions

As it was not possible to categorise the incorrect responses for safeness, there was no kappa analysis associated with this metric to address the safeness aspects of research question 6.

For research questions 3 and 4, and the correctness aspects of research questions 6 and 7, a mixed-model logistic regression analysis was used to model the log odds of being correct for any individual question with terms for level of certainty, test number, year group, number correct being above or below standard on exam and all possible interactions between these factors. The model included random terms for these parameters (with student, except those involving year group as students were only in one year group) and student. The GLIMMIX procedure of SAS 9.4 (SAS Institute Inc., Cary, North Carolina, USA) was used.

For proportion correct with no certainty, compared to chance, the standard error for the proportion was calculated, and from this, confidence intervals calculated.
9.4 Results

9.4.1 General data

Test 1 was sat by 1114 students of whom:

290 were Year 2 (223 were above the standard of 30.8% correct, 67 below standard);
279 were Year 3 (253 were above the standard of 36.6% correct, 26 below standard);
260 were Year 4 (246 were above the standard of 42.5% correct, 14 below standard);
285 were Year 5 (280 were above the standard 47.2% correct, 5 below standard).

Test 2 was sat by 1114 students of whom:

287 were Year 2 (214 were above the standard of 31.2 % correct, 73 below standard);
272 were Year 3 (232 were above the standard of 36.4% correct, 40 below standard);
266 were Year 4 (250 were above the standard of 41.6% correct, 16 below standard);
289 were Year 5 (285 were above the standard of 48.1% correct, 4 below standard).

There was little difference in the number of MCQ options for each level of certainty (median 6, interquartile range 5–8 for all levels certainty).

In total, 51% of questions were answered correctly, 85% (47044/55352) for those with high certainty, 65% (51197/79195) for those with moderate, 38% (47387/125209) for those with low and 32% (22483/60607) for those with no certainty.

Given the greater proportion of candidates with an increased range of ability and number of items for these tests, the reliability of responses, as measured by internal consistency, Cronbach’s α, was higher compared with that in Studies 1-3, and higher still for certainty than for correctness (Table 9.2).
Table 9.2: Internal consistency for responses by year groups for Test 1 and Test 2

<table>
<thead>
<tr>
<th></th>
<th>N candidates</th>
<th>N items</th>
<th>Cronbach's α correctness</th>
<th>Cronbach's α certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1 Year 2</td>
<td>290</td>
<td>150</td>
<td>0.86</td>
<td>0.99</td>
</tr>
<tr>
<td>Test 1 Year 3</td>
<td>279</td>
<td>150</td>
<td>0.80</td>
<td>0.99</td>
</tr>
<tr>
<td>Test 1 Year 4</td>
<td>260</td>
<td>150</td>
<td>0.79</td>
<td>0.98</td>
</tr>
<tr>
<td>Test 1 Year 5</td>
<td>285</td>
<td>150</td>
<td>0.77</td>
<td>0.98</td>
</tr>
<tr>
<td>Test 1 All years</td>
<td>1114</td>
<td>150</td>
<td>0.93</td>
<td>0.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N candidates</th>
<th>N items</th>
<th>Cronbach's α correctness</th>
<th>Cronbach's α certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 2 Year 2</td>
<td>287</td>
<td>150</td>
<td>0.86</td>
<td>0.99</td>
</tr>
<tr>
<td>Test 2 Year 3</td>
<td>272</td>
<td>150</td>
<td>0.83</td>
<td>0.99</td>
</tr>
<tr>
<td>Test 2 Year 4</td>
<td>266</td>
<td>150</td>
<td>0.82</td>
<td>0.99</td>
</tr>
<tr>
<td>Test 2 Year 5</td>
<td>289</td>
<td>150</td>
<td>0.78</td>
<td>0.98</td>
</tr>
<tr>
<td>Test 2 All years</td>
<td>1114</td>
<td>150</td>
<td>0.94</td>
<td>0.99</td>
</tr>
</tbody>
</table>

9.4.2 Results related to research questions

The mixed-model logistic regression analysis was not possible for the four-way interaction due to sparse data, and hence the interaction was not included. In the model with all three-way interactions, certainty–test–standard, certainty–year group–standard, and test–year group–standard interactions were not significant (p=0.30, 0.92, and 0.43 respectively) and so were removed from the model. In the subsequent model test–standard and year group–standard interactions were also not significant (p=0.20 and 0.18 respectively) and were removed from the model.

In the resulting logistic model (Table 9.3), the certainty–test–year group, certainty–standard and test–year group interactions were significant (p<0.0001, p<0.0001 and p=0.001 respectively). The model was used to calculate OR for correctness given certainty, test, standard, and year group (Table 9.3). The ORs were calculated relative to questions answered by Year 2 students who were below standard and had no certainty in their responses.
Table 9.3: Mixed model logistic regression analysis for correct response given factors and interactions between factors

<table>
<thead>
<tr>
<th>Effect</th>
<th>Numerator df</th>
<th>Denominator df</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certainty</td>
<td>3</td>
<td>3446</td>
<td>2098.69</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Test</td>
<td>1</td>
<td>1425</td>
<td>11.84</td>
<td>0.0006</td>
</tr>
<tr>
<td>Year group</td>
<td>3</td>
<td>1139</td>
<td>58.76</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Standard</td>
<td>1</td>
<td>316.5</td>
<td>221.46</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Certainty–test</td>
<td>3</td>
<td>3062</td>
<td>5.39</td>
<td>0.001</td>
</tr>
<tr>
<td>Certainty–year group</td>
<td>9</td>
<td>2730</td>
<td>3.83</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Certainty–standard</td>
<td>3</td>
<td>1999</td>
<td>9.70</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Test–year group</td>
<td>3</td>
<td>1399</td>
<td>5.31</td>
<td>0.001</td>
</tr>
<tr>
<td>Certain–test–year group</td>
<td>9</td>
<td>2908</td>
<td>5.87</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

For all groups, there were significant increases in the odds of being correct with each increasing level of certainty (all p<0.0001, except from no to low certainty, where all p<0.01).

The ORs of answering correctly for low vs. no certainty was 1.15 (95%CI, 1.03–1.29), and higher for those above the standard than those below (p=0.02). For example, for Year 2, the OR for low vs. no certainty, above standard is 2.20/1.38 = 1.59, which is 1.15 times higher than below standard 1.39 (Table 9.4). As the same ratio applies throughout Table 9.4, this can be seen across year groups and levels of certainty. The ORs for moderate vs. low certainty were 1.15 (95%CI, 1.03–1.27), and also higher for those above than below standard (p=0.009). For high certainty vs. moderate certainty the ORs were not significantly different at 1.09 (95%CI, 0.94–1.25, p=0.26).
Table 9.4: Odds Ratios of a correct response relative to below standard students from Year 2 on Test 1

<table>
<thead>
<tr>
<th></th>
<th>Test 1</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 2</td>
<td>Year 3</td>
<td>Year 4</td>
<td>Year 5</td>
</tr>
<tr>
<td>Certainty*</td>
<td>BS</td>
<td>AS</td>
<td>BS</td>
<td>AS</td>
</tr>
<tr>
<td>None</td>
<td>1.00</td>
<td>1.38</td>
<td>1.23</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.39</td>
<td>2.20</td>
<td>1.48</td>
<td>2.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>4.18</td>
<td>7.59</td>
<td>3.95</td>
<td>7.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>13.79</td>
<td>27.17</td>
<td>11.67</td>
<td>23.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Test 2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 2</td>
<td>Year 3</td>
<td>Year 4</td>
<td>Year 5</td>
</tr>
<tr>
<td>Certainty*</td>
<td>BS</td>
<td>AS</td>
<td>BS</td>
<td>AS</td>
</tr>
<tr>
<td>None</td>
<td>1.04</td>
<td>1.44</td>
<td>1.25</td>
<td>1.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.56</td>
<td>2.47</td>
<td>1.68</td>
<td>2.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>4.11</td>
<td>7.45</td>
<td>4.20</td>
<td>7.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>10.41</td>
<td>20.51</td>
<td>13.81</td>
<td>27.21</td>
</tr>
</tbody>
</table>

*BS=below standard; AS=above standard

For each combination of level of ability, year group, and test, there were significant (p<0.0001) differences between levels of certainty.

The score for random selection of options was 16.2% for Test 1 and 15.2% for Test 2. Over the 300 questions, the score for random selection was 15.7%. The percentage correct for all no certainty responses at 32% is higher than chance (p>0.001). Further analysis by test and year group reveal the number correct with no certainty to be higher than chance (p>0.001) (Table 9.5).
Table 9.5: Proportion of no certainty responses that were correct

<table>
<thead>
<tr>
<th>Test</th>
<th>Year group</th>
<th>% of responses (correct responses with no certainty / responses with no certainty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Year 2</td>
<td>26.4% (3449/13052)</td>
</tr>
<tr>
<td>1</td>
<td>Year 3</td>
<td>33.9% (3080/9369)</td>
</tr>
<tr>
<td>1</td>
<td>Year 4</td>
<td>36.6% (2036/5558)</td>
</tr>
<tr>
<td>1</td>
<td>Year 5</td>
<td>41.0% (1083/2639)</td>
</tr>
<tr>
<td>2</td>
<td>Year 2</td>
<td>26.9% (4703/17458)</td>
</tr>
<tr>
<td>2</td>
<td>Year 3</td>
<td>33.4% (3839/11478)</td>
</tr>
<tr>
<td>2</td>
<td>Year 4</td>
<td>37.3% (2985/7990)</td>
</tr>
<tr>
<td>2</td>
<td>Year 5</td>
<td>42.7% (1308/3063)</td>
</tr>
</tbody>
</table>

These proportions increase with year group, aligning with the ORs (Table 9.4). The odds of being correct with no certainty increased with increasing year group and performance standard. Even for the least experienced (Year 2, Test 1), the proportion correct of no certainty responses was higher than expected by random selection.
9.5 Discussion

Initially the results will be summarised in the context of the hypotheses and research questions. The results will be discussed related to feasibility of certainty-in and safeness-of responses, and correctness for levels of certainty. Following this findings not related to research questions will be discussed. Lastly, the strengths and limitations of this study are discussed.

9.5.1 Hypotheses

- Less resource-intensive means for standard-setting and safeness judgements are feasible was confirmed for standard-setting, refuted for safeness judgements
- Undertaking a large scale, high-stakes progress test using certainty-in and safeness-of responses is feasible was confirmed for certainty, with analysis not possible for safeness
- Students of all ability levels, characterised by specified criteria, and experience levels, demonstrate increasing correctness with certainty was confirmed.
- Unsafe responses, as a proportion of incorrect responses, are more likely with increasing certainty for increasing levels of ability was not possible to confirm or refute.
- Students’ ‘no certainty’ responses are more likely to be correct than chance was confirmed

9.5.2 Research questions

1. Are there alternative means for standard-setting and safeness judgements?
   - Yes, for standard-setting.
   - No, for safeness judgements.

2. Can certainty-in and safeness-of responses be included in a large scale high-stakes progress test?
   - Yes, for certainty, two tests of more than 1000 students, answering 150 MCQs.
   - No, for safeness.
3. What is the relationship between response certainty and correctness?
   • The relationship demonstrated increasing correctness with levels of certainty.

4. How do these measures of certainty and correctness vary with year group and ability?
   • All student groups demonstrated increasing correctness with certainty.

5. What is the relationship between response certainty and safeness?
   • Not possible to answer due to sparse data.

6. How do these measures of certainty and safeness vary with year group and ability?
   • Not possible to answer due to sparse data.

7. What is the likelihood of correctness and safeness for responses marked as 'no certainty'?
   • The likelihood of being correct when answering with no certainty was higher than for random selection.
   • Not possible to answer for safeness.

9.5.3 Feasibility of certainty-in and safeness-of responses

The standard-setting proved possible without a significant resource requirement in terms of time from multiple staff members, or of technical expertise and specialised software.

There are significant resource implications of reviewing distractor safeness. Given the complexities of clinical practice and the clinical environment, experienced clinicians are required to make the judgements about distractor safeness. For these tests of 150 questions with a median of six distractors, 900 judgements would be required of each panellist. The prior studies within this thesis highlighted the number of panellists needed was higher than the four, five or six used: increasing the number to ten would mean 9000 judgements would be required.
Alternatives for categorising safety responses need to be considered. Given the basis of certainty-in and safeness-of responses is authenticity to clinical practice, alignment to clinical practice could possibly be used. In practice, the most important incorrect decisions would be those which lead to significant adverse events. The seriousness of an incident can be graded by using the accepted rating system, the Severity Assessment Code (SAC)\textsuperscript{388,389}, although there is a variation in its use\textsuperscript{390}. These codes are used in New Zealand, and healthcare providers are obliged to report SAC1 (resulting in death or permanent severe loss of function) and SAC2 (resulting in permanent major or temporary severe loss of function) events\textsuperscript{391}. A future classification for safety of incorrect responses that could be used, is that if the response would be likely lead to an SAC1 or SAC2 event report should it occur in practice, then it would be classified as potentially unsafe.

The use of certainty of responses in a large scale (students and items) assessment provided feasible. Again, it was possible to obtain assessment information beyond that produced from number correct delivery and marking.

9.5.4 Correctness for given levels of certainty

Yet again, there is the finding that the odds of being correct are strongly associated with the levels of certainty. Furthermore, and unlike Studies 1-3, this study had sufficient power, given the larger number of students and items, to find this even in those students who were least likely to be correct: those in the earliest year group with the least experience whose performance was below standard. Even this group just two months into the course, demonstrated increasing odds of correctness with increasing certainty. This confirms findings in Studies 1-3, which have been discussed along with the relevant literature within the prior study discussion sections of this thesis.

Included in the Year 2 student cohort was a group who were below standard. This subgroup, like the other subgroups, was able to demonstrate increasing correctness with increasing certainty. Could the assessment itself be stimulating behaviour and learning? Formats of assessment that encourage moment-by-moment reflection can stimulate awareness of limitations\textsuperscript{8}. An alternative explanation for the findings is that the cohort and subgroups, even those with minimal experience and below standard performance, demonstrated increasing correctness with certainty because of other factors. These factors may include: the general self-reflection rather
than the specific self-monitoring, a factor specific to these cohorts of students selected into the medical degree, or skills in self-monitoring that had been developed prior to entry into the OMS MBChB degree course.

When comparing those below and above standard, there was a small increase in the OR of being correct between no to low certainty, and low to moderate certainty, but not between moderate to high certainty. Given this small increase in the OR of being correct between levels of certainty for those above, compared with those below standard, it could be that some improvement in self-monitoring occurs for those performing above standard.

The level of correctness for responses with 'no certainty' was higher than chance. Previous studies have found that people underestimate the likelihood of being correct with and select DKO instead, either due to the scoring system and / or their personality 311. Examinees, when forced to answer questions for which they had answered 'don't know' were more likely to be correct than expected by chance 93. Within the self-monitoring literature it was found that responses deferred until a second round were more likely to be correct 8.

This finding that students were more likely to be correct when choosing 'no certainty' than would have been expected by random selection, highlights the balance of safety and efficiency, related to being correct when certain vs. being certain when correct 392. It could be that the students were considering safety ahead of efficiency when they answered the question. As students progress through the course and learn more, it could be that they become more aware of the potential for adverse outcomes, and so become more cautious in their certainty.

Unfortunately, unsafe response categorisation was not possible in this analysis. Higher-than-chance correct responses with no certainty could be acceptable if it reduces the risk of incorrect and unsafe responses being given with higher levels of certainty.

9.5.5 Findings not related to the research questions

The same pattern of internal consistency being higher for certainty than correctness was seen. Unfortunately no data on safeness was possible. This was the first study in this thesis that allowed analysis related to tests with larger cohorts of students (260-
and a larger number of common questions (150). This will be reviewed further in section 10.3.2.

9.5.6 Strengths of this study

The strengths of this study include that this is the first analysis to explore accuracy of correctness using levels of certainty as reflection-in-action for a dataset where there are sufficient student numbers within subgroups defined by experience and performance. Repeating the analysis across two separate tests, and finding the same results, increases the generalisability of the findings.

Acknowledging the need for assistance recognises the role of external cognitive support. The extra information generated from this response format can guide student learning and can guide decisions on student progress. Whilst we have found that student cohorts generally show an indication of accurate self-monitoring, there may well be outlier students who are less able to do this. Identifying and following such students, comparing their results with those from other assessments and evaluating any effect on learning, are areas for future exploration. In particular, the current study reports a snapshot within a single calendar year, but over time response patterns may become apparent.

9.5.7 Limitations of this study

Given the importance of safeness of decisions to clinical decision-making, it is a significant omission that the safety of responses could not be included in this analysis. Safeness is important in the proposed model for how the certainty of a decision is arrived at, and could not be included in this study.
Chapter 10. General discussion and conclusions
10.1 Chapter outline

Firstly, I will summarise the findings across the four studies, then select the main findings and review these in the setting of the literature, whilst acknowledging the potential limitations. Finally, I will draw parallels in self-monitoring clinical decisions and self-monitoring of assessment response selections and suggest implications for the directions for current and future assessment of medical students and future research exploration.
10.2 Summary of findings

Self-monitoring, in a general context, is described as “self-observation and self-control guided by situational cues to social appropriateness”. Self-monitoring as part of daily clinical healthcare practice is reflection-in-action related to clinical-decision-making. Self-monitoring leads clinicians to be aware of when they have reached the limit of their scope of practice. Given the patient and context they are faced with, it is knowing when to ‘look it up’ or ‘defer to others’.

The aim of this thesis was to explore meeting the need for the development of an assessment protocol for self-monitoring that included items based on authentic clinical practice and incorporated certainty-in and safety-of item responses. The plan was to achieve this using appropriate scientific methods and analysis, including psychometric measures, subgroup performance, individual performance factors, and associated performance factors. These analyses would relate to students’ item responses and measures of self-monitoring, which included correctness, certainty, safety (incorrectness, uncertainty and unsafeness) and interactions among these.

The main findings that came from this PhD thesis were:

1. Assessment incorporating response certainty, and to a lesser extent incorrect response safeness, is feasible.
2. The reliability of responses is greatest for certainty, less for correctness and least for safeness.
3. Students of all levels of ability, including low performers, the least experienced, and those below the minimum standard for knowledge, demonstrate increasing correctness with increasing certainty.
4. Unsafe responses occur across all levels of ability, experience and certainty.

These results lead to the conclusion that correctness for given levels of certainty, safeness for given levels of certainty, and correctness of clinical decision-making in assessment responses, are relatively independent outcome measures contributing to information on self-monitoring. Education programmes should consider developing assessments that include certainty-in and safeness-of item responses including feedback related to correctness, certainty and safeness. The categorisation of safeness of responses needs to be further developed.
The implications of findings are that there is a role for response certainty and, with development, unsafeness of responses in MCQ assessment of medical students. Further consideration can be given to extrapolate this to other healthcare professional students and medical practitioners, considering assessment for, of, and as learning.
10.3 Analysis of findings in relationship to prior research literature

10.3.1 Assessment incorporating response certainty, and to a lesser extent incorrect response safety, was feasible

The comparative prior literature is limited when it comes to making inferences regarding self-monitoring for the use of item response certainty and safety, with more related to item certainty alone being used to make inferences on self-monitoring. Given the relative paucity in the use of item level safety, this might corroborate the practical difficulties in making rigorous classifications of this metric. The opinion of a panel of judges has been previously used. Reliability or agreement within the panellists regarding safety judgements among these studies, or others investigating response safeness outside the context of self-monitoring, either do not mention the level of reliability or agreement of the judgements, or highlight it as an area for improvement.

As discussed in Study 4, an alternative means for categorising safety responses needs to be considered. Rather than not, low, moderate, and high risk unsafe incorrect responses, the use of the Severity Assessment Codes (SAC), specifically SAC1 and SAC2 events, as those of greatest risk, could be highlighted. A future response safety categorisation for incorrect responses that could be used is, that if the response would likely lead to a SAC1 or SAC2 event report should it occur in practice, it would be classified as unsafe. Given that there are descriptors of these type of events, and the categorisation could be developed as 'Incorrect no risk of SAC1/2', 'Incorrect and risk of SAC1/2' (incorrect and likely to lead to a near miss for a SAC1/2 event, as would be mitigated within healthcare systems) or 'Incorrect and issue of SAC1/2' (incorrect and likely to lead to a SAC1/2 event). This may improve the rigour, and therefore utility, of using item response safeness.

10.3.2 Reliability of responses is greatest for certainty, less for correctness and least for safeness

Response reliability, as internal consistency using Cronbach’s α, for response correctness, certainty, and safety were calculated for the different assessments, different students, and different stakes to students of these studies. These analyses were remarkably similar, with internal consistency being greatest for certainty, less
for correctness and least for safeness. There is literature, at least for internal consistency of correctness and certainty responses, which reveals similar results.\textsuperscript{133,141,142}

Reliability of responses is influenced predominantly by variability of examinee performance across items with different content (content specificity)\textsuperscript{393} and the range of ability of the students, which affects the range of responses.\textsuperscript{111,387} For the studies in this thesis, the reliabilities of correctness, certainty and safeness of responses were investigated for the same items and same examinees, so spread of ability would not be a factor if a single ‘ability’ factor influenced all correctness, certainty and safeness equally. Given the variation in response reliability, correctness, certainty and safeness of responses, it would not appear to be related to a single factor.

Regarding different content, when there is low reliability, the performance of an examinee on one item is a poor predictor of performance on the next. Therefore, as the influence of content specificity increases, the reliability of responses decreases, as combining different content areas lessens overall item inter-correlations. Therefore, the influence of content specificity is least for certainty, more for correctness and most for safeness. This has been a consistent finding across all the experiments and has added significantly to the prior literature.\textsuperscript{133,141,142}

It has been proposed that certainty is based on a combination of a person’s general propensity for certainty and their specific certainty for a particular question response.\textsuperscript{192} Some people are generally more certain and others more cautious: this variation would then be apparent among people for the same question and level of knowledge.\textsuperscript{213} Factor analytic research has led to a ‘confidence’ factor being postulated as a separate trait.\textsuperscript{142} A person’s self-judgement of ability is more accurate when related to specific detail of content rather than in broader content themes.\textsuperscript{185} A person’s self-judgement of their ability in a broader content theme (e.g. my knowledge of cardiology), as opposed to an item of detail within that theme (e.g. my knowledge of the best drug to treat this cardiac arrhythmia), will be based on more information across the theme, and therefore would be more consistent, but because of true variation of ability within that theme, would more often be incorrect.\textsuperscript{185} People will base their certainty for any given item on item-level certainty, but also on broader levels certainty, up to the potential level of an overall certainty trait. This would explain the finding of prior literature, built upon with this thesis, of internal consistency of certainty responses being higher than internal consistency of
correctness of responses, with the effect on content specificity being greater for correctness of responses.

Although there is no prior literature for comparison, this argument could be extrapolated to internal consistency of response unsafeness. This thesis mirrors previous studies that found the absolute number of unsafe responses chosen to be small. Reliability, such as internal consistency, in measuring such infrequent events has been recognised as an issue. It is almost impossible to get reliable statistical data about rare events; for example, significant motor vehicle accidents are so rare that demonstrating comparisons before and after an intervention with reliability would need extraordinarily large sample sizes. Rare events will have lower statistical reliability, as confirmed across several different studies within this thesis. In addition to this, the classification of the unsafe responses had lower levels of agreement among the reviewing panelists. As discussed above, a classification for potential SAC 1/2 events might improve the classification agreement, but would make the unsafe responses even rarer within the option lists, with even less selection.

If self-judgements reflect general self-concept and the context specific self-efficacy, self-judgements could be reconsidered, such that:

- Level of correctness of response on a given question is informed by a person’s general level of correctness combined with their more task-specific level of correctness, alongside some random variation.

- Level of certainty in the correctness of a response on a given question is informed by a person’s general level of certainty combined with their more task-specific level of certainty, alongside some random variation.

- Level of unsafeness of response on a given question is informed by a person’s general level of incorrectness and unsafeness combined with their more task-specific level of incorrectness and safeness, alongside some random variation.

The internal consistency results would then be explained by level of correctness for a given question being relatively more influenced by task-specific level of correctness; level of certainty for a given question being relatively more influenced by general level of certainty; and the level of safeness for a given question being relatively more influenced by random variation or context factors.
10.3.3 Students at all levels of ability demonstrate increasing correctness with increasing certainty

Repeatedly across the studies in this thesis, the findings revealed that even those who scored lowest, were least experienced, and/or were below the pass standard in terms of numbers correct, had an increasing likelihood of being correct with increasing certainty. A significant body of general literature \(^\text{55,222,340}\) and healthcare literature \(^\text{6,7,345-348}\) has described those of lesser skill as also less aware of their performance, i.e. that the unskilled are unaware. However, when self-judgements reflect self-monitoring, lower performing candidates, like their higher performing peers, do actually reflect in-the-moment, if longer multiple-choice question response time, flagging questions, and changing response are analysed \(^7\). Why is there a discrepancy between the self-monitoring literature and the other self-assessment literature? Can both be correct? Can the unskilled be unaware in self-assessment but be aware in self-monitoring?

Self-monitoring is contrasted to self-assessment, which is a general judgement of one’s ability \(^8\), with self-monitoring linking to the framework of self-efficacy and self-assessment linking to the framework of self-concept \(^5\). The format of assessment may influence results of reflective self-assessment \(^39\). The general literature \(^55,222,340\) and healthcare literature \(^6,7,345-348\) tend to focus on people estimating their rank in a group, estimating their own score or grade, or their numerical probability of being correct. The self-judgement framework, be it self-assessment, self-monitoring or another, and wording of self-judgement criteria are important \(^34\). The parameters of incorrect self-judgement have been postulated \(^140\): overestimation is predicting a performance level higher than actual, also known as functional overconfidence; overplacement is the mistaken belief of being better than others; and overprecision, also known as subjective overconfidence, is being unduly certain in a result which could be higher or lower than that achieved. When these parameters are studied simultaneously, there is variation between these parameters being present or absent \(^340,341\): study participants can overestimate their ability and yet not overplace or be overprecise. The framework, the questions asked and the characteristics of the study population could therefore explain some of the variation in results observed across these studies.
As covered in the literature review section 3.4.4, further explanations of inappropriate self-judgements (e.g. overconfidence) in underperformers are: the 'better than average' effect, regression to an inflated mean, task difficulty, extraction, and ceiling/floor effects. The nature and size of the contribution from these factors remains unresolved. It is accepted that such issues do contribute, but not enough to explain all the original findings. Previously, the finding that 'the unskilled are unaware' was found to be context specific: people know that they are not an elite sportsperson or a musical virtuoso. If a task is too difficult and well beyond a person's scope of ability, they know this. The context-specific task review is akin to self-monitoring. Even the authors of 'the unskilled are unaware' literature accept that the unskilled can self-monitor in certain contexts.

It is proposed that the alignment in this thesis of the wording and format of the certainty questions to self-monitoring in clinical practice explains the difference between the thesis findings and those in the literature. Specifically, the wording referred to seeking advice or to consulting a reference rather than to estimating a peer group ranking or the likelihood of being correct. Response scales work best when they reflect cognitive structuring, which in this case was targeted to reflection-in-action in the moment and seeking assistance. Variations in the format and wording of certainty descriptors can alter meaning, and hence the responses and the inferences that can be drawn from those responses. In other words, the instruction wording in this thesis is more authentic to clinical practice, and this has produced different results.

Where a healthcare professional places in a rank ordering of peers does not relate to their everyday practice. They may not consider it at all, and if forced to do so, having not considered it previously, may not be accurate. It is self-evident that 50% of competent doctors are worse than average (median), and even those above average in terms of performance in daily practice may be below average on any given assessment. It should not be the aim or expectation that, like the fictional town of Lake Wobegon, everyone can be above average. Self-efficacy judgements should relate to criteria rather than norms, with personal success measured against a criterion rather than relative success. This is equivalent to assessment of medical practice, where performance relative to a standard is usually more important than performance relative to peers.
Given variations in format and wording, it is likely that a lower performing individual might be able to self-monitor when this is measured using descriptors of seeking advice or a reference, and be unable to self-assess using descriptors of peer ranking, probability correct or a guess of a final score. Self-monitoring is different to summative self-assessment. Given these findings from the literature and this program of research, how important would the presence of self-monitoring and the relative absence of self-assessment be? This distinction between the specifics of self-monitoring and other more general forms of self-assessment is vital, and this thesis has added to this body of literature highlighting the importance. The self-assessment literature concentrates on accuracy and calibration, whereas self-monitoring, with its roots in self-efficacy literature, concentrates on the consequences of belief. Macro-level self-assessments of competence may obscure minor defects: an assessment that addresses reflection-in-action of daily practice could be more important for ensuring safe and effective clinical performance.

10.3.4 Unsafe responses occurred across all level of student ability

The issues of lower levels of agreement within the judging panel on unsafeness of responses, and the rarity of significant high certainty-in, high unsafeness-of responses does create some potential limitations to this aspect of the thesis. However, a significant number of unsafeness at any level of certainty was still identified, and potentially significant for future practice, as these included those made with high certainty.

This thesis found that even the very lowest performers demonstrated increasing odds of being correct with increasing certainty. However, unsafe responses occurred at all levels of experience and ability, and for some of the studies unsafe responses became more likely for higher certainty responses as ability increased. As the ability of respondents increases from very low, they perceive their increasing ability, so their certainty increases, and at an individual level episodes of incorrect certainty start to appear. A little learning is dangerous thing. It is postulated that the relationship between ability and appropriateness of certainty may be quadratic when those with no ability are more appropriate than those with some ability. Familiarity in one area can lead to assumed ability in another, so knowing a little about several things is more of a problem that knowing a lot about one thing.
As ability increases, a little knowledge in several areas of content could lead not only to an appropriate increase in certainty, but also to misplaced certainty.

Certainty rating is based at least in part on ease of information retrieval. With increasing ability, improving cognitive processing may improve retrieval of information, which usually leads to a correct response, but occasionally leads to an incorrect response, though rarely leads to an unsafe response. Greater knowledge means students recognise more, and may choose an option with high familiarity, even if incorrect, with higher certainty. The use of the 'don’t know' option (DKO) is related to familiarity of the content. People with no familiarity of content, whether due to a lack of knowledge or a lack of retrieval, are more likely to answer that they do not know.

This thesis has replicated some prior findings with a low absolute number of significantly unsafe responses, and similarly, given the small numbers, the unpredictability of such responses. Significant errors are rare and highly content and context specific: they only occur when several specific things align, the Swiss cheese model of medical error, where significant errors only occur when several factors line up (the holes in slices of Swiss cheese). Unexpected, unforeseen, extraordinary circumstances lead to high-certainty high-consequence errors. These events are rare and unexpected, and people accustomed to achieving outcomes will (appropriately) have high levels of certainty for the majority of their decisions.

Medical students are generally high achievers and are used to having high levels of capability for the tasks they face. Self-efficacy for any given task relies on factors including prior performance and self-efficacy for that prior performance. Self-efficacy is influenced by the outcomes of personal experiences. Personal experiences, including success and failure, develop self-efficacy for a task. People learn from the unexpected consequences of their performance that they are aware of, but will not consider outcomes that they have never experienced, even though they may occur.

In Study 1 and Study 2 there was an increase in likelihood of unsafe responses within high certainty incorrect responses, but this was not found in Study 3. There was a different, equally concerning pattern of unsafe higher certainty responses in Study 3, when the responses were not standard, with the correct option
having been removed: students’ higher certainty responses were less likely to be correct and more likely to be unsafe. That Study 3 did not demonstrate the same findings as Study 1 and 2, may reflect the greater question numbers and broader content in Study 3. However the fact that a different concerning pattern of high certainty unsafe responses appeared adds to the benefit of including certainty-in and safeness-of responses in the assessment.
10.4 Potential limitations

10.4.1 Were the students motivated?

The assessment format and timing of some of the studies may have had a negative impact on students, as they may not take research assessments seriously. This could influence responses, and therefore the validity of extrapolations from the results. Students may not concentrate fully on research or non-counting assessments if external summative examinations are considered more important. Perceived consequence influences motivation, and motivation influences performance. Students’ beliefs about the importance of the examination or content affects students’ motivation to prepare for an examination. A lack of effort in preparing for non-consequential tests leads to less motivation to complete the assessment well, and hence lower performance. Examinees consider the severity of the consequence and the chance of those consequences happening. Perceived consequences also alter use of meta-judgements. As with other assessments, any assessment of meta-judgements may get different results dependent on the perceived consequences from the test. The effects are complex: consequence creates motivation and anxiety, and so alters behaviour, but anxiety can cancel the effect of motivation. Although it is interpreted that the effect of motivation on performance is normally distributed, it is possible that there is a normal distribution for a cohort and differing effects for individuals.

The research-based assessments involved volunteer students, who were likely to be more motivated than those who chose not to participate. Although the student interviews in Study 3 confirmed this concern, equivalent results were found across the thesis for different students, across different assessments, with different content, and for different consequences: research only, low-stakes summative, and higher-stakes summative. Although motivation and anxiety may have played a role in responses, given the equivalence of the results, it is likely that the effect was minimal.

10.4.2 Were the students honest?

Students, like all people, will look to maintain their self-image. People explain error by factors other than their own fault, in part to maintain self-image. By ignoring uncertainty and the associated anxiety, self-efficacy is preserved, potentially at the
expense of self-monitoring ». Self-deception, denial and delusion are impediments to self-monitoring ». This paradox was raised as a possibility in the initial literature review, section 2.4.4. Within the healthcare environment people accept responsibility for their actions, but tend to overestimate their contribution to positive outcomes «. outside the healthcare environment people underestimate their contribution to negative outcomes ». Students may be prone to giving socially-desirable responses », or what they believe are socially-desirable responses. The benefit of maintaining self-efficacy is that it adds to performance ». Those who have lower skill levels but higher self-efficacy levels might perform better than those with slightly higher skill level but lower self-efficacy level ». Again, the similar results within this program of research across students and formats implies that it is more likely that cohorts were honest, rather than cohorts were consistently being dishonest. Although potentially limited by representation, the interview participants did not raise dishonest socially-desirable responses as an issue.

10.4.3 Can we extrapolate the results from written assessments to medical student practice and to future clinical practice?

Inappropriate certainty in decision making has been found in both the real world and in laboratory-based research ». However, there is a concern that many investigations of meta-judgements are laboratory-based and therefore may have limited generalisability to the real world ». How does meta-judgement research translate into the workplace »? Laboratory-based performance assessment controls for some, but not all of the irrelevant variance. However, in a workplace, irrelevant variance could be authentic. Easier, less complex tasks are associated with better self-efficacy, and the more complex the task, the weaker the relationship between self-efficacy and performance ».

Assessment of safety often ignores some factors affecting safety, such as other staff input ». Assessment should reflect this authentic practice environment and the social context. Perceived self-efficacy is personal, but people are not socially isolated ». Workplace performance is assessed for individuals who work in teams ». A person’s self-efficacy and performance will be affected by the team and systems in which they work ». Extrapolating from the more controlled environment of medical student assessment to medical student practice and future clinical practice is not just an issue for self-monitoring ». Healthcare professional practice is complex. Trying to make
the written assessments as authentic as possible is one means to minimise irrelevance, and make the results as meaningful as possible. Hence staff and clinicians with content knowledge related to healthcare professional practice, experience of student education, and experience of medical student and post-graduate medical practice in New Zealand, were used to develop items, and standard-set for progress and classification for safeness.
10.5 Implications and future directions

This thesis has implications for medical student assessment, which may in turn inform other healthcare professional student assessment. Extrapolating beyond students, there are potential implications for assessment of trainees and practitioners. In addition to research on the role of other healthcare students, trainees and practitioners, there are other implications for future research.

10.5.1 Healthcare professional clinical decision-making practice and education

Self-regulation is a cornerstone of any profession. Self-regulation has two meanings: at an authority level, it refers to the profession policing its own members, and at an individual level, it refers to self-awareness and control of one’s own practice. Many professions shift responsibility from the authority level to the individual for self-regulation.

The ability to self-assess, 'to know what one knows', is a critical self-regulatory skill. It is expected that medical students will have the ability to self-assess once they graduate. Self-assessment has been found to be underdeveloped during training. The development of self-assessment skills applicable to professional practice should start during undergraduate study. It is important to include development of methods to help medical students develop their self-assessment, and to realise its importance to themselves and patients.

Assessing actual performance and meta-judgements related to performance at the same time is ideal. However, patient and public safety needs to be maintained whilst medical students and doctors are developing their knowledge and self-monitoring. Maintaining authenticity of education, including assessment, to clinical practice whilst maintaining public and patient safety is required. Factors that inform, augment and impair meta-judgements may not be captured if authenticity is reduced, and assessments may fail to extract all the information possible on student ability.

Random guessing in clinical practice, and therefore random guessing by medical students, should not be encouraged. Clinical practice is not dichotomous: it is not 'know or don’t know'. Eva and Regehr suggested that we need to consider whether individuals demonstrate help-seeking behaviour, and to do this we need to incorporate an option for help-seeking behaviour in our assessments.
The framework developed for the thesis is that clinicians will self-monitor within the context of making a decision. They will consider their belief in their capability and their belief in potential consequences to decide how certain they are in those decisions, and whether they need assistance or other resources. The self-monitoring of clinical decision-making includes cognition of the correctness of the decision for levels of certainty and safeness for levels of certainty. Therefore, correctness, certainty and safeness are vital in the consideration of self-monitoring. The similarities between this can be seen in Figure 2.6 and Figure 8.1, both of which have been discussed at more length in the relevant preceding sections. Item level correctness, correctness for levels of certainty, and safeness for levels of certainty, are related to self-monitoring, but are essentially independent.

Clinicians are continually making clinical decisions in their practice. Even a student in a highly supervised environment makes decisions. The student’s decisions might not be the same as those made by practicing clinicians, but they would be decisions, nonetheless. This suggests that education regarding the self-monitoring process might start concurrently with education regarding clinical decision-making. Waiting until graduation to introduce self-monitoring might not be appropriate: the voyage of discovery for personal self-monitoring should not commence in clinical practice. Introducing certainty-in and safeness-of responses in MCQs assessments for medical students has proved feasible across several contexts.

The introduction of self-monitoring to medical students, including certainty-in and safeness-of responses in MCQs assessments, could allow for the identification of students who need additional learning support to improve this ready for commencing practice. Further is the potential to coach and assess students to ensure that they reach a required standard with regard to self-monitoring. This would require a means to define and demonstrate a required standard, which would be an area for further research. To develop research to try to demonstrate that improving self-monitoring and meeting a standard within the context of an assessment as a medical student would lead to improved self-monitoring, and therefore better patient outcomes, as a practicing clinician would be difficult. However, self-monitoring is an important part of clinical practice and like many other facets of clinical practice, demonstrating the relationship to patient outcomes can be difficult.
Given that self-monitoring has been highlighted as part of healthcare professional clinical decision-making practice and education, specifically the practice and education of physiotherapy, occupational therapy, pharmacy, and nursing, there could be a role for certainty-in and safeness-of responses to be considered within these programmes.

10.5.2 Further research

Over the course of the thesis, the use of certainty-in and safeness-of responses have been investigated for cohorts of medical students and individual students, at a single point in time and multiple points in time, and for no stakes, lower stakes and higher stakes assessment. Correctness, correctness for levels of certainty, and safeness for levels of certainty have been demonstrated to change at different rates and in different directions. Correctness, correctness for levels of certainty, and safeness for levels of certainty are important to introduce to medical student education programmes. This needs to be studied further, including across other groups of students and trainees, and across other institutions and countries.

Further exploration of how self-monitoring develops, and how education, assessment and feedback might be developed to optimise the development of this skill, could be studied further.

There are specific areas that will require further exploration that were beyond the scope of this thesis. There is a significant literature on personality and psychological traits and decision-making including certainty and risk taking. The impact of such traits on the development of performance related to clinical decision-making and self-monitoring in practice as well as decision-making and self-monitoring in education environments will add to literature on performance and learning.
10.6 Closing

There is a quote attributed to Mark Twain (1835-1910):

It ain't what you don't know that gets you into trouble.

It's what you know for sure that just ain't so.

This can be paraphrased and developed:

Patient safety is not risked by clinical decisions that you, as a doctor, believe need corroboration: it is the decisions that you erroneously believe do not. You need to learn when your decisions need corroboration.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALM</td>
<td>Advanced Learning in Medicine: Years 4, 5 and 6 of the MBChB degree at the University of Otago Medical School</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>CBA</td>
<td>Certainty based assessment</td>
</tr>
<tr>
<td>CCE</td>
<td>Common component examination</td>
</tr>
<tr>
<td>CFG</td>
<td>Correction for guessing</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>df</td>
<td>Degrees of freedom</td>
</tr>
<tr>
<td>DKO</td>
<td>Don’t know option</td>
</tr>
<tr>
<td>ELM</td>
<td>Early Learning in Medicine: Year 2 and Year 3 of the MBChB degree at the University of Otago Medical School</td>
</tr>
<tr>
<td>EMQ</td>
<td>Extended matching question</td>
</tr>
<tr>
<td>EPE</td>
<td>Early Professional Experience</td>
</tr>
<tr>
<td>F</td>
<td>Fraction</td>
</tr>
<tr>
<td>HCP</td>
<td>Healthcare professional</td>
</tr>
<tr>
<td>LMS</td>
<td>Learning management system</td>
</tr>
<tr>
<td>MBChB</td>
<td>Bachelor of Medicine Bachelor of Surgery (Medical Degree)</td>
</tr>
<tr>
<td>MCCQE</td>
<td>Medical Council of Canada Qualifying Examination</td>
</tr>
<tr>
<td>MCQ</td>
<td>Multiple-choice question</td>
</tr>
<tr>
<td>NS</td>
<td>Not significant</td>
</tr>
<tr>
<td>OSCE</td>
<td>Objective structured clinical examination</td>
</tr>
<tr>
<td>OSPE</td>
<td>Objective structured practical examination</td>
</tr>
<tr>
<td>OMS</td>
<td>Otago Medical School</td>
</tr>
<tr>
<td>OR</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>p</td>
<td>p value</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>( r )</td>
<td>Correlation co-efficient</td>
</tr>
<tr>
<td>RKT</td>
<td>Retained knowledge test</td>
</tr>
<tr>
<td>SAC</td>
<td>Severity Assessment Code</td>
</tr>
<tr>
<td>SAQ</td>
<td>Short-answer question</td>
</tr>
<tr>
<td>SCT</td>
<td>Social cognitive theory</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard error of measurement</td>
</tr>
<tr>
<td>SG</td>
<td>Score group</td>
</tr>
<tr>
<td>VAS</td>
<td>Visual analogue scale</td>
</tr>
</tbody>
</table>
**Associated publications and presentations**

**Chapter 2**

**Peer-reviewed publications**


**Conference proceedings**


**Chapter 3**

**Peer-reviewed publications**


**Chapter 5**

**Peer-reviewed publications**


Conference proceedings


Chapter 6

Peer-reviewed publications

Conference proceedings
M Tweed, T Wilkinson, S Stein, J Smith. Four years of MCQ results reveal consistent patterns in certainty and safety, insightfulness and foresightfulness. Proceeding of the AMEE Conference, Glasgow, Scotland. 2015

Chapter 7

Conference proceedings


Chapter 8

Peer-reviewed publications


Conference proceedings


Chapter 9

Conference proceedings

M Tweed. Medical students are more consistent in their certainty in assessment responses than their correctness of assessment responses. Proceeding of the ANZAHPE Conference, Adelaide, Australia. 2017
References


73. Graber ML. Educational strategies to reduce diagnostic error: can you teach this stuff? Advances in Health Sciences Education. 2009;14(1):63-69.


139. Lichtenstein S, Fischhoff B. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance.* 1977;20(2):159-183.


190. Gardner-Medwin AR, Curtin NA. Certainty-based marking at UCL and Imperial College. 2006.
227. Gardner-Medwin A. Certainty-Based Marking: rewarding good judgment of what is or is not reliable. Paper presented at: Innovation2008; Breckenridge, CO.


257. Mann D. The Relationship between Diagnostic Accuracy and Confidence in Medical Students. American Educational Research Association; 1993; Atlanta.


300. Wilson R, Case S. Extended matching questions: an alternative to multiple-choice or free-response questions. *J of Veterinary Medical Education.* 1993;20(3).


373. Higham PA. Regulating accuracy on university tests with the plurality option. *Learning and Instruction*. 2013;24(0):26-36.
405. Smith ER, Tyler RW. Appraising and recording student progress. 1942.


411. Mattern KD, Burrus J, Shaw E. When both the skilled and unskilled are unaware: Consequences for academic performance. *Self and Identity*. 2010;9(2):129-141.


413. Lingard L. What we see and don’t see when we look at ‘competence’: notes on a god term. *Advances in Health Sciences Education*. 2009;14(5):625-628.


Appendix A.  Environmental scan of good medical practice

These tables documents the results of the environmental scan used, to investigate for the presence of clinical decision-making, seeking further resources and safe practice. For each of these the relevant sections of the various documents is cited.
Table A.1: Clinical-decision-making for consultation, management, treatment options requiring knowledge, synthesised with further information

<table>
<thead>
<tr>
<th>Good Medical Practice, MCNZ</th>
<th>Good Medical Practice, AMC</th>
<th>CANMEDS Physician competency framework</th>
<th>ACGME program requirements</th>
<th>ACGME outcome project</th>
<th>Global minimum essential requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. When you assess, diagnose or treat patients you must provide a good standard of clinical care. This includes: adequately assessing the patient’s condition, taking account of the patient’s history and his or her views, reading the patient’s notes and examining the patient as appropriate; providing or arranging investigations or treatment when needed</td>
<td>2.1 In clinical practice, the care of your patient is your primary concern. Providing good patient care includes: 2.1.1 Assessing the patient, taking into account the history, the patient’s views, and an appropriate physical examination. The history includes relevant psychological, social and cultural aspects.</td>
<td>15 You must provide a good standard of practice and care. If you assess, diagnose or treat patients, you must: a. adequately assess the patient’s conditions, taking account of their history (including the symptoms and psychological, spiritual, social and cultural factors), their views and values; where necessary, examine the patient b. promptly provide or arrange suitable advice, investigations or treatment where necessary</td>
<td>Medical expert. 1.4 Perform appropriately timed clinical assessments with recommendations that are presented in an organized manner 2. Perform a patient-centred clinical assessment and establish a management plan: 2.1 Prioritize issues to be addressed in a patient encounter 2.2 Elicit a history, perform a physical exam, select appropriate investigations, and interpret their results for the purpose of diagnosis and management, disease prevention, and health promotion</td>
<td>IV.A.5.c) Practice-based Learning and Improvement Residents must demonstrate the ability to investigate and evaluate their care of patients, to appraise and assimilate scientific evidence, and to continuously improve patient care based on constant self-evaluation and life-long learning.</td>
<td>Patient care communicate effectively; demonstrate caring and respectful behavior; gather essential and accurate information; make informed decisions about diagnostic and therapeutic interventions, To increase: the likelihood of more appropriate medical decision making and patient satisfaction, the graduate must be able to: listen attentively to elicit and synthesize relevant information about all problems and understanding of their content; synthesize and present information appropriate to the needs of the audience, and discuss achievable and acceptable plans of action that address issues of priority to the individual and community. The graduate must diagnose and manage the care of patients in an effective and efficient way. In order to do so, he/she must be able to: apply basic diagnostic and technical procedures, to analyze and interpret findings, and to define the nature of a problem; The medical graduate should therefore be able to: understand the power and limitations of scientific thinking based on information obtained from different sources in establishing the causation, treatment and prevention of disease; identify, formulate and solve patients’ problems using scientific thinking based on obtained and correlated information from different sources; formulate hypotheses, collect and critically evaluate data, for the solution of problems.</td>
</tr>
</tbody>
</table>
3. In providing care you are expected to: provide effective treatments based on the best available evidence

2.1 In clinical practice, the care of your patient is your primary concern. Providing good patient care includes:

2.1.2 Formulating and implementing a suitable management plan (including arranging investigations and providing treatment and advice).

2.2 Maintaining a high level of medical competence and professional conduct is essential for good patient care. Good medical practice involves:

2.2.6 Providing treatment options based on the best available information.

16 In providing clinical care you must:

a. prescribe drugs or treatment, including repeat prescriptions, only when you have adequate knowledge of the patient’s health and are satisfied that the drugs or treatment serve the patient’s needs

b. provide effective treatments based on the best available evidence

c. check that the care or treatment you provide for each patient is compatible with any other treatments the patient is receiving, including (where possible) self-prescribed over-the-counter medications

Medical expert

2. Perform a patient-centred clinical assessment and establish a management plan:

2.4 Establish a patient-centred management plan

3. Plan and perform procedures and therapies for the purpose of assessment and/or management:

3.1 Determine the most appropriate procedures or therapies

3.3 Prioritize a procedure or therapy, taking into account clinical urgency and available resources

Patient care develop and carry out patient management plans

The graduate must diagnose and manage the care of patients in an effective and efficient way. In order to do so, he/she must be able to: perform appropriate diagnostic and therapeutic strategies with the focus on life-saving procedures and applying principles of best evidence medicine; exercise clinical judgment to establish diagnoses and therapies; manage common medical emergencies;
<table>
<thead>
<tr>
<th>Maintaining practice</th>
<th>Underpinning knowledge applied to practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Medical Practice, MCNZ</td>
<td>Good Medical Practice, AMC</td>
</tr>
<tr>
<td>83. Keep your knowledge and skills up to date throughout your working life: familiarise yourself with relevant guidelines and developments that affect your work; take part regularly in professional development activities that maintain and further develop your competence and performance</td>
<td>7.2 Development of your knowledge, skills and professional behaviour must continue throughout your working life. Good medical practice involves: 7.2.1 Keeping your knowledge and skills up to date. 7.2.2 Participating regularly in activities that maintain and further develop your knowledge, skills and performance. 7.2.3 Ensuring that your practice meets the standards that would be reasonably expected by the public and your peers.</td>
</tr>
<tr>
<td>Recognising and working within limits</td>
<td>Good Medical Practice, MCNZ</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>80. Recognise and work within the limits of your competence.</td>
<td>2.2 Maintaining a high level of medical competence and professional conduct is essential for good patient care. Good medical practice involves: 2.2.1 Recognising and working within the limits of your competence and scope of practice. 2.2.2 Recognising and working within the limits of your competence and scope of practice. 2.2.3 Recognising and working within the limits of your competence and scope of practice. 2.2.4 Recognising and working within the limits of your competence and scope of practice. 2.2.5 Recognising and working within the limits of your competence and scope of practice. 2.2.6 Recognising and working within the limits of your competence and scope of practice. 2.2.7 Recognising and working within the limits of your competence and scope of practice. 2.2.8 Recognising and working within the limits of your competence and scope of practice. 2.2.9 Recognising and working within the limits of your competence and scope of practice. 2.2.10 Recognising and working within the limits of your competence and scope of practice. 2.2.11 Recognising and working within the limits of your competence and scope of practice. 2.2.12 Recognising and working within the limits of your competence and scope of practice. 2.2.13 Recognising and working within the limits of your competence and scope of practice. 2.2.14 Recognising and working within the limits of your competence and scope of practice. 2.2.15 Recognising and working within the limits of your competence and scope of practice. 2.2.16 Recognising and working within the limits of your competence and scope of practice. 2.2.17 Recognising and working within the limits of your competence and scope of practice. 2.2.18 Recognising and working within the limits of your competence and scope of practice. 2.2.19 Recognising and working within the limits of your competence and scope of practice. 2.2.20 Recognising and working within the limits of your competence and scope of practice.</td>
</tr>
<tr>
<td>Referring a patient onto another clinician when limit reached</td>
<td>2. When you assess, diagnose or treat patients you must provide a good standard of clinical care. This includes: taking suitable and prompt action when needed, and referring the patient to another practitioner or service when this is in the patient’s best interests. 2.1 In clinical practice, the care of your patient is your primary concern. Providing good patient care includes: 2.1.4 Referring a patient to another practitioner when this is in the patient’s best interests.</td>
</tr>
<tr>
<td></td>
<td>3. In providing care you are expected to: consult and take advice from colleagues when appropriate</td>
</tr>
<tr>
<td>Consult other clinicians when limit reached</td>
<td></td>
</tr>
</tbody>
</table>

Table A.2: Seeking further resources and/or appropriate referral when limitations are reached
Table A.3: Safe practice and the avoidance of harm

<table>
<thead>
<tr>
<th>Providing safe care minimising risk</th>
<th>Good Medical Practice, MCNZ</th>
<th>Good Medical Practice, AMC</th>
<th>Good Medical Practice, GMC</th>
<th>CANMEDS Physician competency framework</th>
<th>ACCME program requirements</th>
<th>ACCME outcome project</th>
<th>Global minimum essential requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2 Maintaining a high level of medical competence and professional conduct is essential for good patient care. Good medical practice involves: 2.2.2 Ensuring that you have adequate knowledge and skills to provide safe clinical care. 2.2.4 Considering the balance of benefit and harm in all clinical-management decisions</td>
<td></td>
<td></td>
<td></td>
<td>3.3 An important part of the doctor–patient relationship is effective communication. This involves: 3.3.4 Discussing with patients their condition and the available management options, including their potential benefit and harm.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82. Work with patients and colleagues to maintain and improve the quality of your work and promote patient safety. In particular: take part in audit, peer review and continuing medical education</td>
<td>6.2 Good medical practice in relation to risk management involves: 6.2.5 Working in your practice and within systems to reduce error and improve patient safety, and supporting colleagues who raise concerns about patient safety.</td>
<td>22. You must take part in systems of quality assurance and quality improvement to promote patient safety. This includes: a. taking part in regular reviews and audits of your work and that of your team, responding constructively to the outcomes, taking steps to address any problems and carrying out further training where necessary</td>
<td></td>
<td>Medical expert 5. Actively contribute, as an individual and as a member of a team providing care, to the continuous improvement of health care quality and patient safety: 5.1 Recognize and respond to harm from health care delivery, including patient safety incidents 5.2 Adopt strategies that promote patient safety and address human and system factors Professional 2. Demonstrate a commitment to society by recognizing and responding to societal expectations in health care: 2.2 Demonstrate a commitment to patient safety and quality improvement</td>
<td>Residents are expected to: IV.A.5.f. (3) incorporate considerations of cost awareness and risk-benefit analysis in patient and/or population-based care as appropriate;</td>
<td>Systems-based practice incorporate cost awareness and risk-benefit analysis;</td>
<td>Systems-based practice work in interprofessional teams to enhance quality and safety</td>
</tr>
</tbody>
</table>

366
A.1 References

## Appendix B. Scoring schemes

### B.1 Linear/balanced scoring schemes

Possible scores to use for answers and confidence factor

<table>
<thead>
<tr>
<th></th>
<th>Absolutely confident -1.0</th>
<th>Not sure -0.3</th>
<th>Total guess -0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct answer (-1.0)</td>
<td>1.0</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Incorrect answer (1.0)</td>
<td>-1.0</td>
<td>-0.3</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

For this example, the range of results would be as follows:

**Spread of final marks when using the scores above**

<table>
<thead>
<tr>
<th>Absolutely confident, correct answer</th>
<th>Not sure, correct answer</th>
<th>Total guess, correct answer</th>
<th>Total guess, wrong answer</th>
<th>Not sure, wrong answer</th>
<th>Absolutely confident, wrong answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
</tbody>
</table>

**Spread of final marks when scores above are scaled between 0 and 1**

<table>
<thead>
<tr>
<th>Absolutely confident, correct answer</th>
<th>Not sure, correct answer</th>
<th>Total guess, correct answer</th>
<th>Total guess, wrong answer</th>
<th>Not sure, wrong answer</th>
<th>Absolutely confident, wrong answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
</tbody>
</table>

From ¹

*Figure B.1: Balanced scoring system*
Table B.1: Balanced scoring system

<table>
<thead>
<tr>
<th></th>
<th>Certainty of correct</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Certain</td>
<td>Uncertain</td>
</tr>
<tr>
<td>Correct</td>
<td>+4</td>
<td>+1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-4</td>
<td>-1</td>
</tr>
<tr>
<td>Abstain</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

From 1

Table B.2: Balanced scoring system

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Fairly certain</th>
<th>Rational guess</th>
<th>No defensible basis for choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>Omission</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From 2

Table B.3: Balanced scoring system

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0</td>
<td>+0.25</td>
<td>+0.5</td>
<td>+0.75</td>
<td>+1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>-0.25</td>
<td>-0.5</td>
<td>-0.75</td>
<td>-1</td>
</tr>
</tbody>
</table>

From 3

Table B.4: Balanced scoring system

<table>
<thead>
<tr>
<th>Confidence</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>0</td>
<td>+0.5</td>
<td>+1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>-0.5</td>
<td>-1</td>
</tr>
</tbody>
</table>

From 4
**Table B.5: Balanced scoring system**

<table>
<thead>
<tr>
<th></th>
<th>Very confident</th>
<th>Somewhat confident</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>+1.0</td>
<td>+0.5</td>
<td>+0.1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-1.0</td>
<td>-0.5</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

From 1

**Table B.6: Balanced scoring system**

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>+1</td>
<td>+2</td>
<td>+3</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
</tr>
</tbody>
</table>

From 2
B.2 Logarithmic and other alternative scoring systems

![Graphs showing different scoring schemes](image)

From
**Figure B.2: Scoring system**

**Table B.7: Logarithmic scoring system**

UCL confidence-based scoring scheme

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score if correct</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Score if incorrect</td>
<td>0</td>
<td>-2</td>
<td>-5</td>
</tr>
<tr>
<td>Probability correct</td>
<td>&lt;67%</td>
<td>&gt;67%</td>
<td>&gt;80%</td>
</tr>
<tr>
<td>Odds</td>
<td>&lt;2:1</td>
<td>&gt;2:1</td>
<td>&gt;4:1</td>
</tr>
</tbody>
</table>

From [78]

**Figure B.3: Logarithmic scoring system**
Subsequently used by:

Table B.8: Logarithmic scoring system

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>+2</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>+3</td>
<td>-6</td>
</tr>
</tbody>
</table>

From ☐

Table B.9: Logarithmic scoring system

<table>
<thead>
<tr>
<th>Score</th>
<th>Confidence level</th>
<th>Correct</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>+1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>+2</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>+3</td>
<td>-6</td>
<td></td>
</tr>
</tbody>
</table>

From ☐

Table B.10: Logarithmic scoring system

Mapping of the dichotomous response codes onto the polytomous codes

<table>
<thead>
<tr>
<th>Answer</th>
<th>Dichotomous</th>
<th>Confidence</th>
<th>Polytomous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>1</td>
<td>High (C = 3)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid (C = 2)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low (C = 1)</td>
<td>1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>Low (C = 1)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid (C = 2)</td>
<td>-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High (C = 3)</td>
<td>-6</td>
</tr>
</tbody>
</table>

From ☐
Table B.11: Logarithmic scoring system

<table>
<thead>
<tr>
<th>Degree of certainty:</th>
<th>C = 1 (low)</th>
<th>C = 2 (mid)</th>
<th>C = 3 (high)</th>
<th>No reply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark if correct:</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Penalty of wrong:</td>
<td>0</td>
<td>-2</td>
<td>-6</td>
<td>0</td>
</tr>
</tbody>
</table>

From 12,13

Alternative logarithmic scoring systems have included:

Table B.12: Alternative logarithmic scoring system

Marks allocation per question

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very confident</td>
<td>+4 (0)</td>
<td>-2 (-6)</td>
</tr>
<tr>
<td>Fairly confident</td>
<td>+2 (-2)</td>
<td>-1 (-5)</td>
</tr>
<tr>
<td>Not confident</td>
<td>+1 (-3)</td>
<td>0 (-4)</td>
</tr>
</tbody>
</table>

From "

Table B.13: Alternative logarithmic scoring system

<table>
<thead>
<tr>
<th>OLAL TEST 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+4, +2, +1, -2, -1, 0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OLAL TEST 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+3, +2, +1, -3, -2, 0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OLAL TEST 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+3, +2, +1, -3, -2, -1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OLAL TEST 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+5, +3, +1, -3, -2, -1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OLAL TEST 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+5, +3, +1, -2, -1, 0</td>
<td></td>
</tr>
</tbody>
</table>

From "

374
Table B.14: Alternative logarithmic scoring system

Three-level CBM scoring scheme

<table>
<thead>
<tr>
<th></th>
<th>Confidence levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Correct answer</td>
<td>1</td>
</tr>
<tr>
<td>Incorrect answer</td>
<td>0</td>
</tr>
</tbody>
</table>

From

Further scoring systems neither linear, not logarithmic have included:

Table B.15: Alternative scoring system

Scoring matrix for the confidence responses

<table>
<thead>
<tr>
<th></th>
<th>Confidence responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very sure</td>
</tr>
<tr>
<td>Correct</td>
<td>+50</td>
</tr>
<tr>
<td>Wrong</td>
<td>-60</td>
</tr>
</tbody>
</table>

From used by

Table B.16: Alternative scoring system

Scoring matrix for the SACAT answer sheets

<table>
<thead>
<tr>
<th>Answer</th>
<th>Almost guess</th>
<th>Probable guess</th>
<th>Neutral</th>
<th>Fairly certain</th>
<th>Almost certain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>+10</td>
<td>+27</td>
<td>+37</td>
<td>+45</td>
<td>+60</td>
</tr>
<tr>
<td>Wrong</td>
<td>+5</td>
<td>-4</td>
<td>-16</td>
<td>-32</td>
<td>-60</td>
</tr>
</tbody>
</table>

From used by
Table B.17: Alternative scoring system

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>Chance of correct answer</th>
<th>Score for correct answer</th>
<th>Score for incorrect answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 - 25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>26 - 50</td>
<td>+3</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>51 - 75</td>
<td>+4</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>76 - 100</td>
<td>+5</td>
<td>-5</td>
</tr>
</tbody>
</table>

From

Table B.18: Alternative scoring system

<table>
<thead>
<tr>
<th>Confidence in correct</th>
<th>Not at all sure</th>
<th>Fairly sure</th>
<th>Very sure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

From

Table B.19: Alternative scoring system

<table>
<thead>
<tr>
<th></th>
<th>Low confidence</th>
<th>High confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

From

Table B.20: Alternative logarithmic scoring system

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Mod</th>
<th>Low</th>
<th>Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-8</td>
<td>-6</td>
<td>-4</td>
<td>-1</td>
</tr>
</tbody>
</table>

From

Table B.21: Alternative scoring system
Conventional and confidence-weighted scoring schemes

<table>
<thead>
<tr>
<th></th>
<th>Conventional</th>
<th></th>
<th>Confidence-weighted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer</td>
<td>Score</td>
<td>Answer</td>
<td>Confidence</td>
<td>Score</td>
</tr>
<tr>
<td>Right</td>
<td>1</td>
<td>Right</td>
<td>Very sure</td>
<td>4/3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fairly sure</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Guess</td>
<td>2/3</td>
</tr>
<tr>
<td>Wrong</td>
<td>0</td>
<td>Wrong</td>
<td>Guess</td>
<td>1/3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fairly sure</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Very sure</td>
<td>-1/3</td>
</tr>
</tbody>
</table>

From 23

Table B.22: Alternative scoring system

<table>
<thead>
<tr>
<th></th>
<th>Guess</th>
<th>Fairly confident</th>
<th>Very confident</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>+1</td>
<td>+2</td>
<td>+3</td>
</tr>
<tr>
<td>Incorrect low penalty rules</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
</tr>
<tr>
<td>Incorrect high penalty rules</td>
<td>0</td>
<td>-4</td>
<td>-6</td>
</tr>
</tbody>
</table>

From 24

Table B.23: Alternative scoring system

<table>
<thead>
<tr>
<th></th>
<th>Very sure</th>
<th>Fairly sure</th>
<th>Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>+4/3</td>
<td>+1</td>
<td>+2/3</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-1/3</td>
<td>0</td>
<td>+1/3</td>
</tr>
</tbody>
</table>

From 25

Table B.24: Alternative scoring system

<table>
<thead>
<tr>
<th></th>
<th>Sure</th>
<th>Not sure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Incorrect</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

From 26
B.3 References


Appendix C. Ethics committee approvals and research consultation with Māori

The studies undertaking as part of this thesis were approved in accordance with University of Otago policy related to ethical consideration of research and teaching proposals involving human participants.

Study 1  University of Otago Human Ethics Committee reference number D09/20
Study 2  University of Otago Human Ethics Committee reference number D12/379
Study 3  University of Otago Human Ethics Committee reference number D13/120
Study 4  University of Otago Human Ethics Committee reference number D16/037

Research consultation with Māori was carried out for Study 3 and considered by Ngāi Tahu Research Consultation Committee on 21 May 2013 with recommendations and suggestions provided
Kia ora Mike,

We do have a records of this, the reference is: D09/020. It is noted on our database as being received on 9 February 2009. You would not have received an outcome letter at the time. All Cat B applications were considered approved once signed by the HoD.

The entry on our database is minimal and there is no application or any other documentation attached.

Cheers,
Gary
Dear Dr Tweed,

I am again writing to you concerning your proposal entitled “The use of a certainty and safety matrix in summative undergraduate medical student assessment”, Ethics Committee reference number D12/379.

Thank you for your email of 21 June 2017 which provided a copy of the updated/amended Category B Ethics approval. We note that you wish to extend the results to be analysed from 2011-2014, and that statistical analysis will be improved. We confirm the amendments are approved.

In addition the ethical approval for this study has been extended for a further three years, the expiry date is now 27 June 2020.

Your proposal continues to be fully approved. If the nature, consent, location, procedures or personnel of your approved application change, please advise me in writing. I hope all goes well for you with your upcoming research.

Yours sincerely,

[Signature]

Mr Gary Witte  
Manager, Academic Committees  
Tel: 479 8256  
Email: gary.witte@otago.ac.nz

c.c. Professor M Weatherall  
Department of Medicine (Wgnt)
Dear Professor Smith,

I am again writing to you concerning your proposal entitled “The use of certainty and safety response questions in assessment of medical undergraduates: changes in responses over time and rationale for responses”, Ethics Committee reference number 13/120.

Thank you for sending through evidence that consultation is underway with Ngai Tahu Research Consultation Committee; and for confirming that you are consulting with the Division of Health Sciences through Professor Tim Wilkinson.

Thank you for sending through your revised Information Sheet and Consent Form. Please ensure you provide copies of your finalised documents if changes have been made.

On the basis of this response, I am pleased to confirm that the proposal now has full ethical approval to proceed.

Approval is for up to three years from the date of this letter. If this project has not been completed within three years from the date of this letter, re-approval must be requested. If the nature, consent, location, procedures or personnel of your approved application change, please advise me in writing.

Yours sincerely,

Mr Gary Witte
Manager, Academic Committees
Tel: 479 8256
Email: gary.witte@otago.ac.nz

c.c. Professor J K Smith  Associate Dean Research  College of Education
Dear Dr Tweed,

I am writing to confirm for you the status of your proposal entitled “Following individual student’s self-monitoring using progress tests with item level response certainty”, which was originally received on April 23, 2018. The Human Ethics Committee’s reference number for this proposal is D18/127.

The above application was Category B and had therefore been considered within the Department or School. The outcome was subsequently reviewed by the University of Otago Human Ethics Committee. The outcome of that consideration was that the proposal was approved.

Approval is for up to three years from the date of HOD approval. If this project has not been completed within three years of this date, re-approval must be requested. If the nature, consent, location, procedures or personnel of your approved application change, please advise me in writing.

Yours sincerely,

[Signature]

Mr Gary Witte
Manager, Academic Committees
Tel: 479 8256
Email: gary.witte@otago.ac.nz
Tuesday, 21 May 2013.

Professor Jeffrey Smith,
Education Assessment Research Unit - EARU,
DUNEDIN.

Tēnā Koe Professor Jeffrey Smith,

The use of certainty and safety response questions in assessment of medical undergraduates: changes in responses over time and rationale for responses

The Ngāi Tahu Research Consultation Committee (The Committee) met on Tuesday, 21 May 2013 to discuss your research proposition.

By way of introduction, this response from The Committee is provided as part of the Memorandum of Understanding between Te Rūnanga o Ngāi Tahu and the University. In the statement of principles of the memorandum it states “Ngāi Tahu acknowledges that the consultation process outlined in this policy provides no power of veto by Ngāi Tahu to research undertaken at the University of Otago”. As such, this response is not “approval” or “mandate” for the research, rather it is a mandated response from a Ngāi Tahu appointed committee. This process is part of a number of requirements for researchers to undertake and does not cover other issues relating to ethics, including methodology they are separate requirements with other committees, for example the Human Ethics Committee, etc.

Within the context of the Policy for Research Consultation with Māori, the Committee base consultation on that defined by Justice McGeachan:

“Consultation does not mean negotiation or agreement. It means: setting out a proposal not fully decided upon, adequately informing a party about relevant information upon which the proposal is based; listening to what the others have to say with an open mind (in that there is room to be persuaded against the proposal); undertaking that task in a genuine and not cosmetic manner. Reaching a decision that may or may not alter the original proposal.”

The Committee considers the research to be of interest and importance.

The Committee notes and comments that ethnicity data is to be collected as part of the research project and recommends the use of the questions on self-identified ethnicity and descent, these questions are contained in the 2006 census.

The Committee suggests including in the research team a researcher with expertise in analysing and interpreting data by ethnicity.

The Committee suggests dissemination of the research findings to those responsible for curriculum development in the Division of Health Sciences at the University of Otago.

We wish you every success in your research and the Committee also requests a copy of the research findings.
This letter of suggestion, recommendation and advice is current for an 18 month period from Tuesday, 21 May 2013 to 8 November 2014.

Nāihaku noa, nā

Mark Brunton
Kaitukangaere Rangahau Māori
Research Manager Māori
Research Division
Te Whare Wānanga o Ōtāgo
Ph: +64 3 479 8738
Email: mark.brunton@otago.ac.nz
Web: www.otago.ac.nz
Appendix D. Writing good MCQs
### Table D.1: Writing good MCQs

<table>
<thead>
<tr>
<th>Advice for writing good questions</th>
<th>Qualities of good questions</th>
<th>Flaws in poor questions</th>
<th>Plan for this PhD thesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Design each item to measure an important learning outcome</td>
<td>Test important material</td>
<td>Item content will be taken from curriculum map and/or learning outcomes, and verified by clinicians</td>
</tr>
<tr>
<td>Format</td>
<td>Use an efficient item format – response style</td>
<td></td>
<td>Formats will be pen/paper or computer delivery where feasible</td>
</tr>
<tr>
<td>Items</td>
<td>Make sure all items independent</td>
<td></td>
<td>Items will be selected to ensure no items link to each other</td>
</tr>
<tr>
<td>Lead-in</td>
<td>Use of vague terms such as frequently, occasionally or rarely (among others)</td>
<td>Lead-ins will avoid vague semi-quantitative terms</td>
<td></td>
</tr>
<tr>
<td>Lead-in</td>
<td>Use of EXCEPT as part of the question formulation</td>
<td>Lead-ins will avoid words such as ‘except’</td>
<td></td>
</tr>
<tr>
<td>Lead-in</td>
<td>Failure to pass the Hand Cover Test (HCT) increases uncertainty about the question being asked, or leaves the examinee guessing.</td>
<td>Lead-in will be written such that the stem provides vital information to get correct answer ( ), it should not be possible to look at lead-in and options, and not the stem, and get correct answer ( ).</td>
<td></td>
</tr>
<tr>
<td>Lead-in</td>
<td>Use of negative(s) in the question.</td>
<td>Questions will avoid negative words, but include negative concepts such as side-effects and contra-indications</td>
<td></td>
</tr>
<tr>
<td>Options</td>
<td>Advice for writing good questions</td>
<td>Qualities of good questions</td>
<td>Flaws in poor questions</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Options</td>
<td>Follow normal rules of grammar</td>
<td>Be similar in grammar</td>
<td>Grammatical clues, found when using sentence completions. The option with an incorrect grammatical flow is automatically eliminated by most candidates</td>
</tr>
<tr>
<td>Options</td>
<td>Make sure the correct answer is correct or clearly best</td>
<td>Stem should lead to only one possible answer</td>
<td>Logical clues, based on information in the stem also being used in the correct keyed option. Words repeat, where the stem has a complete or part of a word that is clearly identified in the correct keyed option. Convergence cues, usually based on multiple facts used in the options.</td>
</tr>
<tr>
<td>Options</td>
<td>Avoid verbal clues that might enable students to select correct or eliminate incorrect answers</td>
<td>Be similar in content</td>
<td></td>
</tr>
<tr>
<td>Options</td>
<td>Put least wording as possible in the answers</td>
<td>Distractors should be short and to the point</td>
<td>Clinicians will verify the authenticity of wording of the options</td>
</tr>
<tr>
<td>Options</td>
<td>Advice for writing good questions</td>
<td>Qualities of good questions</td>
<td>Flaws in poor questions</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------</td>
<td>-----------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Options</td>
<td>Make distractors plausible and attractive to uninformed</td>
<td>Be similar in length</td>
<td>The use of absolute terms such as never, always, only, which are seldom appropriate qualifiers for clinical statements. Inaccurate information, including implausible options.</td>
</tr>
<tr>
<td>Options</td>
<td>Vary length of correct response to eliminate length as a clue</td>
<td>Avoid ‘all of above’ or ‘none of above’</td>
<td>The longest option is the correct keyed option because of the number of qualifying statements added to justify it as the best choice.</td>
</tr>
<tr>
<td>Options</td>
<td>Avoid all of the above none of the above</td>
<td>Avoid ‘all of above’ or ‘none of above’ (NOTA or AOTA)</td>
<td>Options will avoid ‘all’ or ‘none’ response types</td>
</tr>
<tr>
<td>Options</td>
<td>Vary correct answer random</td>
<td></td>
<td>Options will be placed in alphabetical or numerical order, to make responses easy to find in long lists</td>
</tr>
<tr>
<td>Options</td>
<td>Control for difficulty by changing stem or responses</td>
<td></td>
<td>Clinicians will verify the authenticity of the stems and/or options</td>
</tr>
<tr>
<td>Advice for writing good questions</td>
<td>Qualities of good questions</td>
<td>Flaws in poor questions</td>
<td>Plan for this PhD thesis</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-----------------------------</td>
<td>-------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Stem</td>
<td>Present a single clearly formulated problem in the stem of each item</td>
<td>Stems should be focused</td>
<td>Unclear language, ambiguities, gratuitous information, vignette not required etc. Use of interpreted data.</td>
</tr>
<tr>
<td>Stem</td>
<td>State the stem of the item in simple clear language</td>
<td>Stems should be simple</td>
<td>Lost sequence in presentation of data, failure to use ranges and mixed units, as well as overlapping data, or no normal values given.</td>
</tr>
<tr>
<td>Stem</td>
<td>Put as much wording as possible in the stem</td>
<td>Stem should contain majority of information.</td>
<td></td>
</tr>
<tr>
<td>Stem</td>
<td>State the stem of the item in the positive form, wherever possible</td>
<td>Stems should be positively phrased</td>
<td></td>
</tr>
<tr>
<td>Stem</td>
<td>Emphasize negative wording whenever it is used in the stem of an item</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
D.1 References

Appendix E. Core presentations

The Otago Medical School MBChB Course Curriculum Map includes Core elements (Core Presentations, Core conditions, and Core Professional Activities).

Related to each of these is a series of learning outcomes:

1. A medical complication of pregnancy (Gestational hypertension, diabetes, or anaemia)
2. A woman with pelvic pain / discomfort (acute or chronic)
3. Abdominal distension, visceromegaly / mass / ascites
4. Abnormal vaginal bleeding (excluding pregnancy)
5. Abnormal vaginal discharge
6. Acute abdominal, loin or groin pain
7. Acute fracture, dislocation, or joint / muscle / tendon injury
8. Acute or chronic back or neck pain.
9. Acute or chronic mono-articular joint pain, swelling, stiffness, locking, or giving way,
10. Acute or chronic poly-articular joint pain, swelling, or stiffness.
11. Acute urinary retention
12. Anaemia / pallor
13. Anxiety
14. Arterial blood gases - abnormal acid-base balance (acidosis or alkalosis)
15. Blunt or penetrating injury around the eye
16. Breast lump or other changes
17. Bruising / purpura / petechiae
18. Burns
19. Chest pain
20. Chronic movement disorder (ataxia, bradykinesia, choreoathetosis, clumsiness, dystonia, myoclonus, psychogenic disorder, spasticity, tic, or tremor).
21. Chronic shortness of breath, wheeze, or cough
22. Chronic skin rash
23. Chronic / intermittent abdominal pain
24. Coughing up blood (Haemoptysis).
25. Delirium (acute confusional state / acute mental change)
26. Developmental delay, intellectual disability, dysmorphic features, communication disorders, or learning difficulties
27. Diabetes - advice on ongoing self-management
29. Disturbance of bowel habit (constipation, diarrhoea, faecal incontinence, flatulence, pain, steatorrhoea).
30. Dizziness or vertigo
31. Ear ache or discharging ear
32. Failure to thrive.
33. Febrile illness. Including night sweats or rigors (especially meningitis, pneumonia, bone, joint and soft tissue sepsis, septicaemia), and hyperthermia
34. Feeding problems - mother and infant
35. Gastro-intestinal bleeding (haematemesis or melaena
36. Haematuria/haemoglobinuria
37. Hallucinations and/or abnormal beliefs
38. Head injury
39. Head pain (headache)/oro-facial pain
40. Hearing loss/ tinnitus (acute or chronic)
41. Heart murmur.
42. Hoarseness / voice change.
43. Hypertension (including hypertensive emergency)
44. Immune deficiency immunosuppression or abnormalities of White Blood Cells
45. Impaired consciousness
46. Interpret and explain blood test results (FBC, ESR, CRP, LFTs, amylase, U&E, calcium, phosphate, lipids, glucose, serological anti-body profiles (viral hepatitis, autoantibodies, gastrointestinal infections))
47. Jaundice.
48. Localized or generalised lymphadenopathy
49. Lower gastro-intestinal/rectal bleeding
50. Lower -urinary tract symptoms - male or female- (nocturia, frequency, urgency, hesitancy, dysuria, haematuria, reduced flow, dribbling or incontinence)
51. Male sexual dysfunction
52. Malnutrition or eating disorder
53. Memory loss/ forgetfulness
54. Menopausal symptoms
55. Multiple injuries
56. Nausea and/or vomiting
57. Neck lump (especially Goitre + abnormal thyroid function)
58. Normal pregnancy and puerperium
59. Numbness or paraesthesia’s
60. Obstetric emergency
61. Obstructed airway (cyanosis, choking, paroxysmal coughing, stridor)
62. Onset of swelling in the calf or leg
63. Pain in calf, thigh, or buttocks on walking (intermittent claudication)
64. Palpitations/abnormal heart rhythm
65. Peripheral oedema
66. Physical, sexual, psychological or emotional abuse within and outside the family group, including elder abuse
67. Pigmented skin lesion or pigment patch
68. Polyuria/polydipsia
69. Problems in the puerperium
70. Proteinuria
71. Recent onset shortness of breath
72. Recurrent falls
73.  Red eye
74.  Request for contraception or sterilisation
75.  Request for help with infertility
76.  Risk-taking behaviours such as alcohol and other drug use, sexual activity and violence.
77.  Scrotal/groin pain or swelling
78.  Self-abuse or suicidal thoughts/ behaviour
79.  Shock (hypovolaemic, cardiogenic, or distributive)
80.  Skin or subcutaneous lump(s)
81.  Skin ulcer(s)
82.  Sleeping problems.
83.  Slow or accelerated growth
84.  Splenomegaly
85.  Sudden or gradual loss of vision.
86.  Symptoms of depression (lowering of mood, reduction of energy, and decrease in activity)
87.  Transient/ episodic alteration/ loss of consciousness (syncope or seizure)
88.  Vaginal bleeding in pregnancy
89.  Weakness, either focal or generalised.
90.  Weight loss, anorexia, cachexia