EVALUATING PROFESSIONALISM, TEAMWORK AND LEADERSHIP IN MEDICAL UNDERGRADUATES

Thesis submitted in accordance with the requirements of
the University of Liverpool

For the degree of
Doctor of Medicine

By

Michael John Moneypenny
BSc(Hons) MBChB(Hons) FRCA FHEA

February 2015

Department of Medical Education
Institute of Psychology, Health and Society
Faculty of Health & Life Sciences
University of Liverpool

Dedication

This work is dedicated to my partner, Catherine. Without her understanding, patience and love it would not have been completed.

Acknowledgments

I would like to acknowledge my supervisors, Professor Helen O'Sullivan and Professor Arpan Guha for their long-term encouragement, support and advice. I would also like to acknowledge Dr Simon Mercer and Dr Kate Glennon, who were co-raters on the assessment tool part of the study.

EVALUATING PROFESSIONALISM, TEAMWORK AND LEADERSHIP IN MEDICAL UNDERGRADUATES

DECLARATION

The work contained in this thesis was carried out during my tenure as a clinical research fellow at the Centre for Excellence in Developing Professionalism, School of Medical Education, University of Liverpool. The thesis is based on original studies performed by me. The local ethics committees granted approval for all the studies.

I certify that the work contained in this thesis is my own. It has not been submitted for any other degree or other qualification.

Michael John Moneypenny

Candidate

Table of contents

Abstract

The complexity of healthcare is increasing due to new discoveries in the treatment of disease, the multiple pathologies of an ageing population and changes in working patterns and job roles. In addition, an increase in professional, regulatory and public scrutiny has led to revelations of poor care leading to preventable disability and death. Inquiries into sub-standard care have uncovered a number of professional lapses, in particular failures in teamwork and leadership.

Medical undergraduates are future doctors. Their ability to work effectively within teams and to lead when necessary will therefore have a significant impact on the health of the population. In order to improve leadership and teamwork abilities we must be able to assess them. A literature review searching for a tool to assess teamwork and leadership in the medical undergraduate was carried out. As a consequence of an unsuccessful search, a tool was developed and evaluated, using data from existing tools and from a series of focus groups with medical undergraduates. The focus groups and an examination of the reasoning of assessment participants also informed a study on the justifications for failing to challenge poor performance by a more senior member of staff.

The tool data showed adequate validity and reliability for formative assessments in a simulated environment. The focus groups and examination of reasoning highlighted the continued existence of the medical hierarchy, with steep authority gradients.

This tool can be used in formative assessments, but further research is required before it is used outside the simulated environment and consideration must be given to psychometrics, feasibility and cost. The teaching and assessment of teamwork and leadership, should be given more time in the undergraduate curriculum and medical schools, regulatory bodies, deaneries and trusts should collaborate on minimising the unprofessional behaviours of senior healthcare personnel.

Preface

On the 15th January 2009, US Airways flight 1549 struck a flock of Canada geese soon after take-off from New York's LaGuardia airport. With the loss of both engines, the Airbus A320-200 was turned into a 70 tonne glider. Through the actions of Captain Chesley "Sully" Sullenberger and his crew, the plane ditched safely on the Hudson river approximately 3 minutes after the bird strike. Lauded for his leadership and calmness under extreme stress, Sullenberger told a crowd at his hometown welcoming that: "…I know I can speak for the entire crew when I tell you we were simply doing the job we were trained to do." (Associated Press, 2009)

Four years earlier, on the 29th March 2005, Elaine Bromiley, a 37-year-old mother of two, was scheduled to undergo a routine sinus operation under general anaesthesia. Unfortunately there were complications with managing her airway after she had been anaesthetised. Two consultant anaesthetists and a consultant ENT surgeon were unable to obtain a definitive airway and she suffered hypoxic brain damage. Her life support was switched off some days later. An Independent Report into her death criticised the lack of communication within the team (Harmer, 2007). Her husband, Martin Bromiley, an airline pilot and expert in human factors training in aviation, stated:

> "The lead anaesthetist… in his own words 'lost control'. There was a question mark, in the inquest, about who people felt was in charge at different points… There was certainly a breakdown in the decision-making processes and it would appear that the communication processes dried up amongst the consultants." (Clinical Human Factors Group, 2008)

The aim of this MD project was to use or develop a tool, which would allow for the assessment of teamwork and leadership in medical undergraduates. The project was funded by the Centre for Excellence in Developing Professionalism (CEDP) at the University of Liverpool's School of Medical Education. CEDP's focus on professionalism meant that the project explored undergraduate

teamwork and leadership through a “professional practice” lens. Teamwork and leadership in this context are seen to be desirable attributes in their own right but also represent observable, external, manifestations of an unseen, internal, professional character.

Overview of Chapters

Each chapter has its own introduction, however an overview of the chapters will provide a précis of the MD.

Chapter 1: Introduction

The Introduction provides the background to the MD in terms of the development of the concept of professionalism within medical education, as well as the notion that teamwork and leadership may be seen as components of professionalism. The chapter concludes with a justification of the need for assessment and the rationale for using simulation-based assessment.

Chapter 2: Literature review

This chapter details a literature review of the various databases in order to scope out the existing (to end of July 2009) assessment tools. Unfortunately we were unable to retrieve a suitable assessment tool for leadership and teamwork of the individual medical undergraduate. We therefore decided to develop our own tool with input from the literature review, focus groups and assessment tool methodology literature.

Chapter 3: Focus group study

This chapter details a number of focus groups carried out with 4th year medical students. Wear and Kuczewski (2004) state: "the theory of professionalism should be constructed from a dialogue with those we are educating" (p.2) and Duffield and Spencer (2002) argue that acceptability of an assessment tool requires input from those who are going to be assessed. Therefore, the students' views on professionalism, how it has changed over the years, what behaviour is expected of them as medical students, and the barriers to professional behaviour, were explored. The focus groups also allowed the medical students to provide input into the assessment tool by discussing their notions around the qualities of a good or bad leader and a good or bad teamworker.

Chapter 4: Development and evaluation of the assessment tool

This chapter details the development of the leadership and teamwork assessment tool, using information gathered from the preceding work and additional review of definitions of teamwork and leadership, as well as assessment tool methodology. The evaluation of the assessment tool is presented and generalisations are considered.

Chapter 5: Challenge the leader

A concept which was strongly supported in the focus groups was the importance of challenging poor performance by other team members. In addition, the failure to speak up appropriately has resulted in catastrophic failures both within and outside medicine. The inclusion of two challenge points allowed the assessors to evaluate performance against the tool. The lack of challenge by a number of the participants was felt to be worthwhile of further study and their rationales for not challenging or delaying their challenge is explored in this chapter.

Chapter 6: Conclusion

The conclusion reviews the findings of the MD, integrates them within the current research base and makes suggestions for further study. It also includes a list of recommendations based on a reflection of the MD findings.

CHAPTER 1: INTRODUCTION

Professionalism

Until the early 1980s, there was no mention of the concepts of profession or professionalism within medical education (Arnold, 2002). Medical schools' primary focus was on the teaching of scientific knowledge rather than on the development of a humanistic or professional doctor (Cruess and Cruess, 1997). Behaviours that are now classified as elements of professionalism were instead considered to be non-cognitive skills (Keck et al., 1979). In 1981 the American Board of Internal Medicine (ABIM) appointed a Subcommittee on Evaluation of Humanistic Qualities of the Internist. Arising from a need to be assured that candidates who were certified by the Board had "satisfactory interpersonal and communicative skills" as well as "humanistic qualities", the Subcommittee defined the essential humanistic qualities as "integrity, respect, and compassion" (Krevans and Benson, 1983). The ABIM supported additional research into the definition and assessment of these qualities and, along with the American College of Physicians Foundation and the European Federation of Internal Medicine, co-authored "Medical Professionalism in the New Millennium: A Physician Charter" (American Board of Internal Medicine, 2002). The Charter combined the humanistic qualities with the values of a profession (Arnold and Stern, 2006), resulting in 3 fundamental principles and 10 professional responsibilities (Table 1-1)

Table 1-1: Principles and responsibilites of the physician (ABIM, 2002)

Fundamental Principles
• Primacy of patient welfare • Patient autonomy • Social justice
Professional Responsibilities
• Commitment to professional competence • Commitment to honesty with patients • Commitment to patient confidentiality • Commitment to maintaining appropriate relations with patients • Commitment to improving quality of care

• Commitment to improving access to care • Commitment to a just distribution of finite resources • Commitment to scientific knowledge • Commitment to maintaining trust by managing conflicts of interest • Commitment to professional responsibilities

Developments in the UK paralleled those in the United States. In 1983, Parliament passed the Medical Act (Medical Act, 1983), which defined the modern role of the General Medical Council (GMC), to "protect, promote and maintain the health and safety of the public" (General Medical Council, 2014). In 1995, the GMC published "Good Medical Practice", which detailed, for the first time, the duties and responsibilities of doctors and defined the principles of good medical practice (General Medical Council, 1995). In 1997, the president of the GMC called for a new agreement between medicine and society in order to maintain effective medical professionalism (Irvine, 1997).

In 1998 the Secretary of State for Health established an inquiry into the care of children who underwent cardiac surgery at Bristol Royal Infirmary. The final report made a number of recommendations, including a need for the education, training and continuing professional development of healthcare professionals in teamwork and leadership (Department of Health, 2001).

In 2005, the Royal College of Physicians' Working Party on Medical Professionalism published a definition of medical professionalism and a set of 6 commitments which doctors should uphold (Tallis, 2006) (Table 1-2)

Table 1-2: Definition and commitments of professionalism (Tallis, 2006)

Definition of Professionalism
Medical professionalism signifies a set of values, behaviours, and relationships that underpins the trust the public has in doctors.
Commitments
• Integrity

- Compassion
- Altruism
- Continuous improvement
- Excellence
- Working in partnership with members of the wider healthcare team

There is no universally accepted definition of professionalism (Birden et al., 2014), and the need for such a concept has been challenged (Erde, 2008, Hodges et al., 2011). This view has been supported by Cruess et al. (2010) who argue that "professionalism" will differ between countries and cultures, as it is based on a social contract between medicine and society. For example, research carried out by Ho et al. (2012) found that Taiwanese medical students were more influenced by Confucian relationalism than by the principles of "Western" professionalism. Chandratilake (2014) argues for a middle-ground, stating that "there is a core area of professionalism that extends not only across cultures, but also across disciplines even as certain elements of professionalism are 'context'- specific" (p.345).

Birden et al. (2014) argue that the major conceptual divide in professionalism is between seeing it as a set of attributes and seeing it as an overarching ethos. This divide may also be seen in the educational milieu between the need to train and assess either individual characteristics or overall character (Whitehead et al., 2013).

Additional research over the past few years has shown that unprofessional behaviour has adverse effects beyond those already discussed in the introduction and focus group chapters. Unprofessional behaviour has also been found to result in poorer patient outcomes (Patel et al., 2011), reduced patient satisfaction (Bahaziq and Crosby, 2011, van Mook et al., 2012), increased recruitment costs (Rosenstein, 2011) and reduced employee satisfaction (Reiter et al., 2012)

The concept of lapses of professionalism, as developed by Ginsburg et al. (2000) and Stern (2006), which avoids labeling a person as "professional" or "unprofessional" but rather looks at behaviour in context has gained additional following (O'Flynn et al., 2014). Wong and Trollope-Kumar (2014) argue that "contemporary constructivist theories of identity formation conceive identity to be a dynamic phenomenon that is continually negotiated and co-constructed within a social and relational environment" (p.490). This means that "professional identity (is) a multidimensional, evolving and lifelong process…" (p.490).

Teamwork and Leadership

Whether as components of clinical or professional competence, effective leadership and teamwork are increasingly recognised as essential skills (Nutter and Whitcomb, 2001, Frankel et al., 2006, Darzi, 2008, Salas et al., 2009).

In the UK, the National Confidential Enquiry into Maternal Deaths stated that poor teamwork was a leading cause of substandard obstetric care (Cooper and McClure, 2005). In the US, the Institute of Medicine's landmark report "To Err is Human: Building a Safer Health System" calculated that medical error was the eighth most common cause of death (Kohn et al., 2000). Its follow-up report "Crossing the Quality Chasm" emphasised the need for improved leadership and teamwork in clinical practice (Chakraborti et al., 2008).

This view is supported by research in trauma resuscitation and simulation, which has detailed the pivotal role played by a competent leader (Holzman et al., 1995, Hoff et al., 1997, Cooper and Wakelam, 1999, Flin and Maran, 2004, Hjortdahl et al., 2009). Effective leadership improves team performance and goal achievement (Helmreich, 1997, Hamman, 2004, Marsch et al., 2004). Research has shown that good teamwork reduces errors (Morey et al., 2002, McCulloch et al., 2009), reduces mortality and morbidity rates (Buelow et al., 2008, Neily et al., 2010) and improves patient safety (Lingard et al., 2004, Baker et al., 2005a). From a social perspective, as the population ages, more patients will present with multiple health problems, requiring effective interdisciplinary teamwork and leadership (Hall and Weaver, 2001, Xyrichis and Lowton, 2008).

Leadership and teamwork has received international consideration with the development of the CanMEDS framework by the Royal College of Physicians and Surgeons of Canada (Frank, 2005), which has mandated "leadership" as a core competency. "Teamwork and leadership" was also one of the nine content areas addressed by the US Health Resources and Services Administration's Undergraduate Medical Education for the 21st Century (UME-21) project (O'Connell and Pascoe, 2004).

In 2013, the Francis report detailed the failings in care at the Mid-Staffordshire NHS trust (Francis, 2013). Poor leadership, by nursing, medical and boardroom staff was highlighted as a particular area of concern. It also called for "effective teamwork between all the different disciplines and services" (p.110). The Francis report also emphasised the importance of good leadership: "The common culture and values of the NHS must be applied at all levels of the organization, but of particular importance is the example set by leaders" (p.78). The Francis report was followed by the Keogh Mortality Review (Keogh, 2013) which reported on 14 hospitals with high standardised mortality ratios. Poor leadership was again identified as a cause of patient harm. The Prime Minister then asked Don Berwick, former president of the US Institute for Healthcare Improvement, to produce a report entitled "A promise to learn - a commitment to act: Improving the Safety of Patients in England" (National Advisory Group on the Safety of Patients in England, 2013). Recommendations included: "All NHS leaders and managers should actively address poor teamwork" (p.16) as well as guidance on the shift in leadership behaviours required. In 2014, the Vale of Leven Hospital Inquiry Report stated: "Poor leadership also contributed to an inadequate standard of nursing care" (p.11) and has an entire section entitled "Failures in leadership" (Lord MacLean, 2014).

Worldwide there has been an increase in defined leadership curricula and the provision of training in teamwork and leadership (O'Sullivan and McKimm, 2011c). In 2012, the GMC published "Leadership and management for all doctors" (General Medical Council, 2012) and in 2013, the GMC updated its guidance to doctors of the standards that are expected of them (General Medical Council, 2013). These documents made it clear that effective teamworking and leadership is a professional obligation, expected of all doctors. In the UK, the Medical Leadership Competency Framework (MLCF) developed by the NHS Institute for Innovation and Improvement and Academy of Medical Royal Colleges (2010) provided a blueprint for mapping competencies. The BMJ and the Open University have developed a Clinical Leadership Programme with courses in Clinical Leadership. The NHS Leadership Academy (2013) has developed a Healthcare Leadership Model, whose aim is the professionalization

of leadership at all levels of healthcare. In 2011 the Faculty of Medical Leadership and Management was established in order to "promote the advancement of medical leadership, management and quality improvement at all stages of the medical career" (FMLM, 2014).

In addition to failures in teamwork and leadership, the changes in workload and working arrangements for doctors have increased the risks of poor leadership and teamwork. An increase in the workload of most doctors (van Mook et al., 2009b) has co-incided with the introduction of the European Working Time Directive (NHS Employers, 2009) and the New Deal for Junior Doctors which has resulted in a significant decrease in working hours (Royal College of Physicians, 2012). This has led to an increase in handovers and the need to ensure that care is maintained despite frequent changes within the care team. Poor handovers have been shown to be a major cause of teamwork breakdowns resulting in medical error (Singh et al., 2007). These regulations have therefore accentuated the need for effective teamwork and leadership

Teamwork and Leadership in the Undergraduate

Although "Good Medical Practice" may have been the first time the duties and responsibilities of a doctor were defined in the UK, it was preceded by a GMC publication aimed at undergraduate education, "Tomorrow's Doctors", in 1993 (General Medical Council, 1993). In section 40.3, the publication lists a number of attitudinal objectives expected of the undergraduate including:

> "40.3 (h) awareness of personal limitations, a willingness to seek help when necessary, and ability to work effectively as a member of a team" (p.15)

It could therefore be argued that the standards expected of the undergraduate medical student predate those of the postgraduate doctor. In addition, the publication of "Good Medical Practice" led to a move away from the traditional curriculum. The focus of assessment shifted from process and structure to outcomes (Carraccio et al., 2002) and the curriculum needed to include the teaching and assessment of professionalism by examining medical student behaviour and attitudes (Fowell et al., 2000).

In the updated edition of "Tomorrow's Doctors" the GMC for the first time explicitly confirmed the need for leadership education (General Medical Council, 2009). In a section entitled "Overarching outcome for graduates", Tomorrow's Doctors states:

> "…graduates will make the care of patients their first concern… using their ability to provide leadership and to analyse complex and uncertain situations" (p.14) (General Medical Council, 2009)

In the same document, the GMC stresses the need for doctors to undertake leadership roles and be able to accept being led by others. Table 1-3 details the leadership and teamworking competencies detailed in the 2009 "Tomorrow's doctors" and "Medical students: professional values and fitness to practice".

Table 1-3: Undergraduate leadership and teamwork competencies

Tomorrow's Doctors (2009)
• Using their ability to provide leadership (7) • Effective communication and teamworking (14 (j)) • Demonstrate ability to build team capacity and positive working relationships and undertake various team roles including leadership and the ability to accept leadership by others (22)
Medical students: professional values and fitness to practice (2009)
• Integrity • Compassion • Altruism • Continuous improvement • Excellence • Working in partnership with members of the wider healthcare team

The philosopher John Locke said: "I have always thought the actions of men the best interpreters of their thoughts". By requiring medical undergraduates to display good leadership and teamwork the GMC may, along with Stern and Ginsburg (2004), be supporting the idea that the behaviours of individuals reflect their underlying beliefs and attitudes.

While good leadership and teamwork in both under- and post-graduate medicine may seem to be a pressing need, it is less clear how "good" leadership and teamwork can be assessed. This is vital because as Cohen (2006) states: "If it can't be measured, it can't be improved" (p.613).

Assessment of Teamwork and Leadership

Assessment serves a number of purposes. It allows the assessors to prove to stakeholders (regulatory bodies, the public, etc.) that certain benchmarks have been reached, provides data for programme evaluation, provides feedback to learners and is one of the most important drivers for learning (Newble and Jaeger, 1983, Fowell et al., 2000).

As well as being the regulatory body for medical professionals, the GMC decides whether a medical school is entitled to issue medical degrees (Medical Act 1983). It carries out inspection visits and issues quality assurance reports about each UK medical school. The GMC expects medical schools, through outcome-based education, to provide students with the opportunities to develop their skills to a high standard (Brown and Doshi, 2006). It also expects medical schools to develop and use appropriate tools and processes to ensure these standards have been met. Other regulatory and educational bodies have added their own thoughts to the need for clinical leadership and teamworking (CanMEDS, UME-21) and the importance of medical leadership in effecting change has been detailed elsewhere (O'Sullivan and McKimm, 2011a). There is therefore a duty placed on the medical school to teach and assess the desired characteristics of the future doctor. As Pawlina et al. (2006) stated:

> "modern group practice organisations require a physician to be not only a member of a team, but also a leader, often of several teams that must work together… Thus, in order to be successful in today's healthcare system, graduating physicians must possess new knowledge and competencies such as professionalism, leadership, and teamwork skills" (p.609)

van Mook et al. (2009c) explained that decisions need to be made about the number and type of assessors, as well as the location and frequency of assessment. As Ginsburg et al. (2000) lamented:

> "Knowledge and skills are rigorously evaluated by written and oral exams, standardized patient scenarios, and ward evaluations. However, evaluation of behaviors, including professionalism, is often implicit, un-systematic and, therefore, inadequate" (p.S6)

This inadequate teaching and evaluation leads to undergraduates who are unprepared for teamwork and leadership (McNair, 2005, Rudland and Mires, 2005). This view is supported by O'Connell and Pascoe (2004) who, in their article "Undergraduate Medical Education for the 21st Century: Leadership and Teamwork" state: "Further efforts to demonstrate the mastery of new skills in this important content area… are needed" (p.S51). The Ottawa 2010 conference produced a consensus statement and recommendations on "performance in assessment" which highlighted the outstanding issue of "ensuring all aspects of competence are assessed, including 'softer' competences of leadership, professionalism etc." (p.371) (Boursicot et al., 2011)

Assessment can also be used to provide feedback to a candidate, highlighting areas of good and poor performance. (Rowntree, 1987). Cohen (2006) states: "they don't respect what you expect; they respect what you inspect" (p.613). Therefore, defining the desirable behaviours which are to be demonstrated by a good teamworker or leader, and assessing them, allows the undergraduate to appreciate what the other stakeholders consider to be important. Assessment may also result in an increase in the effort that students apply (van Mook et al., 2009a) and encourage desirable changes in their future behaviours (Norcini et al., 2011).

The desire to assess individual performance informed the entire MD project. The clinical teams that currently form within the acute care setting are often *ad hoc* (Leach et al., 2009). This is one of the distinctions between teams in aviation (where assessment also focuses on the individual pilot (Flin et al., 2003)) and medicine, as opposed to teams in industry and the military (Flin and Maran, 2004). This transient nature of healthcare teams supports an argument that the unit of assessment should be the individual team members rather than the team

itself (Murray and Foster, 2000). Additionally, focusing on the team as a whole may preclude specific feedback (Wright et al., 2009) and may lead to blame-allocation and avoidance of ownership of identified team weaknesses. However, as Lingard (2009) argued, competent individuals may form incompetent teams. Therefore, the use of simulation allowed for the assessment of the individual within a team, as opposed to looking solely at the individual (Hodges, 2013, Roberts, 2013). The literature review therefore discarded tools which examined teams rather than people and the focus groups and tool development referred to the behaviour of the individual.

Using Simulation to Assess Teamwork and Leadership

There are a number of methods of assessment including:

- Written (e.g. multiple choice questions, single best answers (SBA), essays),
- Oral (e.g. *viva voce*)
- Self-asssessment
- Observation-based (e.g. Objective Structured Clinical Examination (OSCE), Clinical Evaluation Exercise (MiniCEX), Direct Observation of Procedural Skill (DOPS), Objective Structured Assessment of Technical Skills (OSATS))

Written and oral assessment

The assessment tool should match the domain being examined. For example, SBAs are thought to be good tests of theoretical knowledge and reasoning skills. Although knowledge of teamwork and leadership could be assessed using a paper-based or oral exercise, knowing what to do and doing it are very different skills (Boulet et al., 2003). As Hawkins et al. (2009) state:

> "Knowledge and attitudes, while indicative of the effectiveness of educational experiences, do not necessarily predict subsequent demonstration of effective skills or behaviours or patient care outcomes. Because the performance in the domain of professionalism may be influenced as much by personal characteristics and social context as by knowledge, the link between knowledge and performance in practice may well be weaker in this domain than in the area of clinical practice" (p.352)

This concept is supported by a paper by Rodgers et al. (2010) which showed that written evaluation does not predict clinical performance in advanced cardiac life support (ACLS).

Self-assessment

Doctors seem to have limited ability to assess themselves accurately (Claridge et al., 2003, Davis et al., 2006, Jones et al., 2008, Eva and Regehr, 2011). In particular, doctors who are least skilled are least able to carry out self-assessment (Edwards et al., 2003, Davis et al., 2006). However, Sargeant et al. (2010) and Plant et al. (2013) found that self-assessment may be robust with the proviso that it is supported by qualitative feedback and a personal tutor. In a study by Weller et al. (2013), intensive care teams were reliably, in comparison with external assessors, able to self-assess performance in terms of ranking, but scored themselves significantly higher than the assessors. In addition, Eva and Regehr (2011) found that, although self-assessment is a poor measure of competence, self-monitoring ("a moment-by-moment awareness of the likelihood that one maintains the skill/knowledge to act in a particular situation" (p.311)) is positively correlated with performance.

Observation-based assessment

Although the practice of teamwork and leadership may best be demonstrated during clinical practice, this environment is beset with difficulties. Some leadership and teamwork skills are only displayed, and tested, in crises; accurately evaluating these skills would require one to wait for a crisis and then have an assessor available. Additionally, a number of studies have shown the biases which affect assessments in clinical practice, including the halo effect (Rowland-Morin et al., 1991, Paisley et al., 2005), gender bias (Wang-Cheng et al., 1995) and relying on indirect evidence of performance (Mazor et al., 2008).

It was therefore decided to assess teamworking and leadership using the "Shows How" stage of Miller's learning pyramid (Miller, 1990) (Fig.1-1), in a context which is as realistic as possible (Murray and Foster, 2000, van Mook et al., 2009a). According to van der Vleuten and Schuwirth (2010) this involves using hands-on patient (standardized)

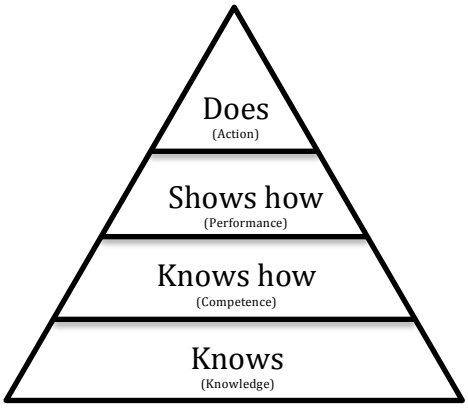


Figure 1-1: Miller's assessment pyramid (1990)

scenarios or simulation as the stimulus and direct observation, checklists or rating scales as the response.

Few assessment tools, as Epstein and Hundert (2002) argue, allow us to observe candidates in real-life situations, however the use of simulation allows us to create a realistic scenario (Ker et al., 2006). The need for teamwork training to occur in a realistic setting has been emphasised, Barrow (2012) refers to a "complex, sociological space", while Sharma et al. (2011) discuss the benefits of "sociological fidelity" in interprofessional simulated learning. Additional benefits of simulation include safety, reproducibility and audio-visual recording (Gaba et al., 1998, Maran and Glavin, 2003, Gaba, 2004, Issenberg et al., 2005, Rall and Gaba, 2005), focused feedback (Kneebone et al., 2002) and the experience of critical or rare events (Hofmann, 2009).

Epstein (2007) states: "High-technology simulation is seen increasingly as an important learning aid and may prove to be useful in the assessment of knowledge, clinical reasoning, and teamwork" (p.392). This concept is supported by a consensus statement on the criteria for good assessment from the Ottawa 2010 conference (Norcini et al., 2011), which states: "Research done over the past few decades is very supportive of the use of [simulation] in assessment…" (p.209) In addition, in order to be able to reliably rate a teamwork or leadership behaviour one requires a dynamic, interactive context as provided by high-fidelity simulation (Wright et al., 2009) which can replicate the stressors found in real-life (Driskell and Johnston, 1998). In their paper "Assessment methods in medical education", Norcini and McKinley (2007) state: "simulation is very realistic and provides an excellent assessment of skills that are difficult to obtain in any other fashion" (p.243).

In addition, simulation has been used extensively both to train (Leonard et al., 2004, Okuda et al., 2009, Østergaard et al., 2004) and assess (Gaba et al., 1998, Wallin et al., 2007) teamwork and leadership behaviours. The University of Dundee has developed a postgraduate ward simulation exercise which assesses teamwork and leadership skills such as the "ability to prioritise competing

demands, make safe informed decisions, prescribe safely and manage the care of three patients" (Stirling et al., 2012). In their focus group study with medical undergraduates, Paskins and Peile (2010) found that students thought the use of mannequin-based simulation allowed them "to develop teamwork skills not only as a more efficient team member but also as a leader" (p.572). This finding supports our use of simulation to assess teamwork and leadership. The authors also found that students exposed to simulation were more confident in their clinical attachments and that they valued both repeated exposure and the feedback on their performance.

Simulation-based training and formative assessment is becoming routine across a range of healthcare disciplines (Gaba et al., 2001, Gaba, 2004). The UK's Nursing and Midwifery Council, for example, allows up to 300 hours of simulation-based training out of a total of 2300 hours of clinical apprenticeship (Nursing & Midwifery Council, 2007). Khan et al. (2011) argue for the use of simulation in the longitudinal assessment of performance, helping "to bridge the gap between the classrooms and the clinical environments". An assessment tool which is applicable in a simulation setting may therefore be of some use.

Lastly, Stern (2006) described the characteristics of an effective assessment as one that:

1) Occurs in as realistic an environment as possible,
2) Includes some form of conflict of values or beliefs and
3) Shows the reasoning behind the actions rather than merely the "correct" response

The third characteristic is supported by Ginsburg et al. (2004) and Rees and Knight (2007). Use of a simulated scenario, which required teamwork, leadership and included a poorly performing "leader", followed by a think-aloud allowed us to meet all three criteria in a controlled, replicable manner.

CHAPTER 2: LITERATURE REVIEW

Introduction

In a quote in their 2005 paper, Reed et al. (2005) summed up the difficulties inherent in carrying out a systematic review of the medical education literature:

> "(Identification of relevant sources and execution of a comprehensive search strategy) are uniquely challenging to reviewers of educational interventions because no single database is devoted to medical education" (p.1080)

However, despite these difficulties, the systematic review remains a pre-requisite for any detailed study. By sifting and filtering the extant literature the systematic review informs us of the current knowledge base, prevents us from re-inventing the (educational) wheel and directs us to areas of potential and utility.

This systematic review aimed to answer the following, linked research questions:

- What tools have been described for measuring teamwork and leadership in individual nurses and/or physicians, both under- and/or post-graduate?
- What are the psychometric properties of these tools?
- What are the practicalities of tool deployment?
- Have any tools been shown to change performance?

The literature databases were selected based on a recommendation from Reed et al. (2005) and included all the databases chosen by Jha et al. (2007) in their systematic review of studies assessing and facilitating attitudes towards professionalism. The search strategy was designed to be inclusive and was reviewed by the MD supervisors.

The analysis was carried out by the MD student with referral to the MD supervisors at key stages to ensure a robust and cogent study.

As the focus of the MD was the assessment of teamwork and leadership in medical undergraduates, the literature review was conducted in order to ascertain whether or not a tool already existed which could be used to carry out our assessments.

Methods

Design

A systematic review method was used based on guidance from the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPICC)(2007). EPPICC has been developing methods for systematic reviews since 1993 (EPPIC, 2009) and it has a strong track record in reviews of educational practice (Odom et al., 2005).

Sample

Inclusion criteria consisted of studies published between the beginning of the given database and the end of July 2009, in English and relating to humans. The cut-off point of July 2009 was used because the simulation-based assessments (see Chapter 4) were scheduled to take place in September/October of 2009. Studies were excluded which did not describe an assessment tool (e.g. review, editorial), did not describe an evaluation of an assessment tool or were not used in healthcare workers.

Search strategy

The following electronic databases were searched from the day they were launched until end of July 2009.

- Pubmed
- Scopus
- EBSCOHost (Cumulative Index to Nursing and Allied Health Literature (CINAHL), PsycINFO, Educational Resources Information Centre (ERIC))
- Web of Science (Conference Proceedings Citation Index, Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index)

The following searches were carried out:

- Pubmed
    - Teamwork* OR Leader* OR Non-technical AND Nurs* OR Medic* AND Undergraduate OR Postgraduate OR Trainee* OR Junior* OR

Student* AND Assess* OR Evaluat* OR Measur* OR Judg* OR Rating
    - In article title or abstract
    - From 1966 to 31st July 2009
- Scopus
    - Teamwork* OR Leader* OR Non-technical AND Nurs* OR Medic* AND Undergraduate OR Postgraduate OR Trainee* OR Junior* OR Student* AND Assess* OR Evaluat* OR Measur* OR Judg* OR Rating
    - In article title, abstract or keywords
    - From 1960 to 31st July 2009
- EBSCOHost
    - Teamwork* OR Leader* OR Non-technical AND Nurs* OR Medic* AND Undergraduate OR Postgraduate OR Trainee* OR Junior* OR Student* AND Assess* OR Evaluat* OR Measur* OR Judg* OR Rating
    - In article title or abstract
    - Before August 2009
    - Limiters
        - ERIC: English
        - CINAHL: English, Human
        - PsycINFO: English, Human
- Web of Science
    - Teamwork* OR Leader* OR Non-technical AND Nurs* OR Medic* AND Undergraduate OR Postgraduate OR Trainee* OR Junior* OR Student* AND Assess* OR Evaluat* OR Measur* OR Judg* OR Rating
    - In article title, abstract, keywords or author keywords
    - Before August 2009
    - Limiters: English

Hand-searching was carried out on the reference lists of reviews of assessment of teamwork and/or leadership (Fletcher et al., 2002, Baker et al., 2005b, Chakraborti et al., 2008) and on the references listed in the original retained articles.

Materials

A data extraction form, based on the data collection recommended by EPPICC (2007), was developed (see Appendix 2-1). This form allowed for detailing of study characteristics such as the journal in which it was published, country where study took place, type of study, sampling method used (Table 2-1), number of participants, etc. The form also requested data based on what constitutes a good assessment tool (Quality Assurance Agency, 2006, van der Vleuten and Schuwirth, 2006): tool psychometrics (validity, reliability) and tool practicalities (acceptability, educational impact, and feasibility).

Table 2-4: Sampling method of assessment tools (Teddlie and Yu, 2007)

Sampling Method	Description
Randomised	Able to determine the non-zero probability of inclusion of every member of the population of interest
Purposeful	Non-random selection of members of a population of interest
Convenience	Non-random selection of "captive" or volunteer members who are easily accessible

Tool psychometrics

Validity

Validity refers to "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p.1) (Messick, 1991).

There are two commonly used classification systems for detailing evidence supporting test validity. According to the Standards of Educational and Psychological Measurement produced by the American Educational Research Association and others (1999), all validity is construct validity (Downing, 2003). This means that construct validity is the over-arching term and five other types

of validity (content, response process, internal structure, relationship to other variables, consequences) help to support or refute construct validity.

The traditional method for assessing validity is to provide three types of validity evidence:

- Construct
    - Convergent: The assessment provides similar scores on related items (internally) and the score agrees with other tests which measure the same variable (externally)
    - Divergent: The assessment provides different scores on unrelated items both internally and externally
- Content
    - Representation: The assessment is a valid representation of a given theoretical construct
    - Face: The assessment appears to measure what it has been designed to do.
- Criterion
    - Concurrent: The assessment correlates with performance on a different assessment performed on the same day
    - Predictive: The assessment correlates with performance on a different assessment performed on some day in the future

The decision to use one or other classification method is primarily user preference and, for the purposes of this study, the traditional method was used.

Reliability

Reliability refers to the ability of those using the tool to achieve reproducible scores (van der Vleuten and Schuwirth, 2005). The most common types of reliability evidence provided is "inter-rater" (raters agree with each other) and "intra-rater" (raters agree with their own score, usually after repeating the scoring at a later date). There are a number of statistical analyses used, details of these are provided in the Discussion section below.

Tool practicalities

Acceptability

The acceptability of an assessment tool may refer to a number of groups, including those being assessed and those assessing, as well as interested parties such as regulatory bodies and the public (General Medical Council, 2011). The assessment tools were analysed for the provision of acceptability evidence for any of the above groups.

Educational impact

This term refers to the influence of the assessment tool on those being assessed. For example, the content, the format and the timing of the assessment may have differing effects (Schuwirth and van der Vleuten, 2010). The data extraction form assessed educational impact via reference to the Kirkpatrick (1998) training criteria (Table 2-2).

Table 2-5: Kirkpatrick's (1998) 4 levels of training criteria

Kirkpatrick Level	Description
1	Reaction e.g. How did participants feel?
2	Learning e.g. Has participant performance improved immediately post-assessment?
3	Behaviour e.g. Has participant performance improved long-term (3-6 months)?
4	Results e.g. Has participant performance improvement led to other benefits such as fewer complaints?

Feasibility

Feasibility is "the degree to which the assessment method selected is affordable and efficient for the testing purpose" (Norcini and McKinley, 2007), i.e. cost-effectiveness. The number of assessors and assessments, as well as the infrastructure required to carry out the assessments impact on the feasibility of the tool.

Procedure

After duplicate articles had been removed, the article titles were assessed for possible relevance to the review and irrelevant articles were excluded. The abstracts for all articles that were obviously or possibly relevant were read. Irrelevant articles were again excluded. The complete articles pertaining to all obviously or possibly relevant abstracts were obtained from library services and read. Irrelevant articles were excluded and all relevant articles had data extracted. The selection process was reviewed and approved by Dr O'Sullivan.

Analysis

Study characteristics/demographics were detailed in frequency tables. Heterogeneity of retained studies precluded the use of statistical analysis and therefore a narrative analysis was detailed.

Results

Study characteristics

Number of studies

The results for the search procedure are detailed in Appendix 2-2. The initial search strategy yielded 4130 references, of which 1687 were duplicates. The remaining 2434 references were assessed by title and the 2328 non-relevant references rejected. The remaining 106 references were assessed by abstract and the 75 non-relevant references were rejected. The remaining 31 references were read in full. 15 of these were rejected because they were either not looking at an individual or were not describing a tool. The remaining 16 articles had data extracted.

The references for these 16 articles were hand-searched, along with the references of reviews, which resulted in an additional 580 references. 501 of these references were rejected by title. Of the remaining 79 references, 33 were duplicates already found in the previous search strategy. Therefore 46 article abstracts were assessed and 37 articles were rejected at this stage. 9 articles were read in full and 2 articles were rejected because they were not describing a tool or were describing a tool which had not been evaluated. The remaining 7 articles had data extracted.

There were therefore 23 articles (16 from the original search and 7 from the hand search) which had data extracted. The references for the 23 articles are listed in Appendix 2-3.

Journals, subject areas and country of study

The articles were published in a variety of journals. The journals, numbers and study numbers are detailed in Table 2-3.

Table 2-3: Articles by journal of publication

Journal	Number (Study numbers)
Annals of surgery	2 (13, 14)
Critical Care Medicine	2 (10, 11)

Medical Education	2 (5, 20)
Medical teacher	2 (9, 21)
Pediatrics	2 (1, 2)
Academic Emergency Medicine	1 (3)
Annals-Academy of Medicine Singapore	1 (17)
British Journal of Anaesthesia	1 (8)
Journal of Interprofessional Care	1 (18)
Learning in Health & Social Care	1 (15)
Medical Care	1 (6)
Medical Education Online	1 (16)
Resuscitation	1 (4)
Simulation in Healthcare	1 (22)
Surgical endoscopy	1 (12)
Teaching and Learning in Medicine	1 (7)
The American Journal of Surgery	1 (19)
World Journal of Surgery	1 (23)

The journals were classified according to subject area as follows (Table 2-4)

Table 2-4: Journals according to subject area

Subject area	Number (Study numbers)
Medical Education	7 (5, 7, 9, 15, 16, 20, 21)
Surgery	5 (12, 13, 14, 19, 23)
Critical Care/Resuscitation	3 (4 , 10, 11)
Medicine	2 (6, 17)
Pediatrics	2 (1, 2)
Anaesthesia	1 (8)
Emergency Care	1 (3)
Interprofessional Care	1 (18)
Simulation	1 (22)

The articles described research carried out in four countries (Table 2-5)

Table 2-5: Publications by country

Country	Number (Study numbers)
USA	10 (1, 2, 5, 6, 7, 10, 17, 18, 21, 22)
UK	9 (4, 8, 12, 13, 14, 15, 16, 19, 23)
Canada	3 (3, 9, 11)
New Zealand	1 (20)

Samples

The majority of studies used uniprofessional (medical) subjects, the remainder used multiprofessional subjects (Table 2-6).

Table 2-6: Publications by number of professions

Professions (Type)	Number (Study numbers)
Uniprofessional (Medical)	19 (1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 13, 14, 16, 17, 19, 20, 21, 22, 23)
Multiprofessional (Medical, Nursing)	3 (10,12,15)
Multiprofessional (Medical, Nursing, Social Care)	1 (18)

The majority of studies used postgraduate subjects, the remainder used undergraduate subjects or both (Table 2-7).

Table 2-7: Publications by graduate status of participants

Graduate status of subjects	Number (Study numbers)
Postgraduate	14 (1, 2, 4, 6, 8, 9, 10, 11, 12, 13, 14, 16, 19, 23)
Undergraduate	8 (3, 5, 7, 15, 17, 18, 20, 21)
Both	1 (22)

A variety of recruitment strategies were used (Table 2-8).

Table 2-8: Publications by recruitment strategy

Recruitment strategy	Number (Study numbers)
Convenience sampling	10 (1, 6, 9, 10, 11, 12, 15, 17, 20, 22)
Purposeful sampling	6 (2, 3, 4, 5, 7, 16)
Not specified	5 (13, 14, 18, 19, 21)
Not applicable (scripted scenarios)	2 (8, 23)

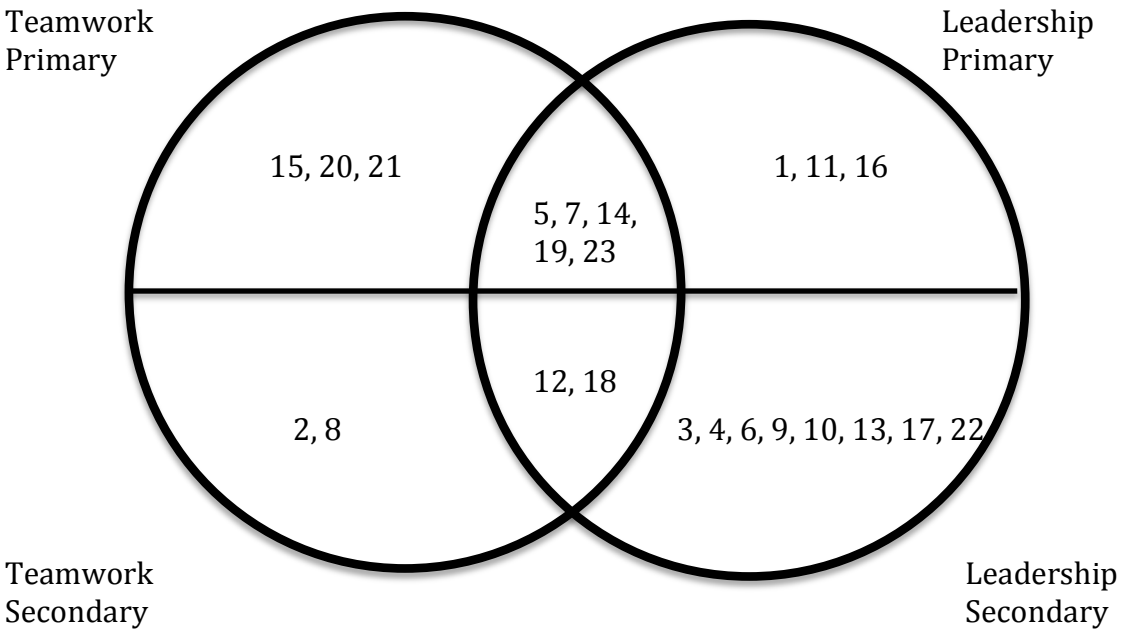
Sample size, where specified, varied from 134 to 6. Appendix 2-4 details the sample sizes in terms of profession and graduation status.

Aims and objectives

2 studies had the primary aim of evaluating an existing assessment tool and 9 had the primary aim of developing and evaluating a new assessment tool. Of these 11 studies, 3 assessment tools assessed individual leadership only, 3 assessed individual teamwork only and 5 assessed both.

The remaining 12 studies provided data on an assessment tool in order to support their primary aim. Of these 12 studies, 8 assessed individual leadership only, 2 assessed individual teamwork only and 2 assessed both (Figure 2-1).

Figure 2-1: Publications by primary assessment focus



Study designs

A number of study designs were used including: pilot studies (14), surveys (4), pre- and post-interventional (3) and observational (1). One study used data gathered at baseline from a RCT and was classified as "other".

The studies took place either in a simulated environment (14) or collected data from the workplace/university environment (9).

16 of the studies used a short intervention such as a simulated scenario in order to assess teamwork and/or leadership. 7 studies reported longer assessment timeframes (weeks or months).

7 studies used peer assessors. In this instance, peer was defined as a person who is at the same stage of training (either under- or post-graduate). In 2 studies the assessor(s) were not identified and in the remaining 14 studies non-peer assessors were used.

A breakdown of the study design criteria is provided in Appendix 2-5.

Assessment tool development and tool types

Tool development

A variety of methods were used in tool development, ranging from author preference to literature review and large-scale interviews. Broad categories of tool development are provided in Table 2-9.

Table 2-9: Publications by tool development method

Tool development method	Number (Study numbers)
Existing tool (+/- modifications)	11 (4, 11, 12, 13, 14, 15, 17, 19, 20, 21, 22)
Author preference	3 (3, 5, 10)
Existing guidelines	2(2, 18)
Literature review, interviews	2 (8, 23)
Existing guidelines and modified Delphi	1 (1)
Existing guidelines and needs assessment	1 (9)

survey	
Review of curriculum and expert opinion	1 (16)
Unknown	2 (6, 7)

Tool type

Of the 22 studies which described the tool type, 19 used a Likert scale (ranging from 4 to 9 points) and 3 used a checklist (performed/not performed/ (borderline)). All the checklists assessed leadership only, with the number of actions to be observed ranging from 9 to 30. The Likert scales used Likert items of teamwork ranging from 1 to 5 and items of leadership ranging from 1 to 27.

Figure 2-2: A 5-point Likert scale using 3 Likert items

Tool psychometrics

Validity

The 23 articles were analysed for evidence of validity of the given tool using both the traditional and more recent definitions of validity. 19 studies provided validity evidence. The most common type of validity evidence offered was construct validity (14), followed by content (10), and criterion (2). The raw data is provided in Appendix 2-6.

Reliability

The 23 articles were analysed for evidence of reliability of the given tool. 16 articles provided reliability evidence. The most common type of reliability evidence provided referred to inter-rater reliability (13), followed by internal consistency (8), and intra-rater reliability (1). 6 of the studies discussed rater standardisation or calibration.

The most common statistic provided for inter-rater reliability was intra-class correlation (4), followed by Cronbach's α (2), Cohen's K (2), r_{wg} (2), Pearson

correlation coefficient (1) and percentage agreement (1). In one study inter-rater reliability was said to be good but the test was not specified. Internal reliability was tested using Cronbach's alpha in all 9 studies which provided data. In the single study which discussed intra-rater reliability (SN: 11) the test was not specified.

The differing measures of reliability provide an agreement score. In the same way that a P value <0.05 may be considered statistically significant, so the score may be considered to show differing degrees of reliability, from "none" to "absolute". The raw data is provided in Appendix 2-7.

Studies whose primary aim was the development and/or evaluation of an assessment tool were more likely to present psychometric data. (Table 2-10)

Table2-10: Publications by study aim and reliability/validity data

	Primary Aim: Number (%)	Secondary Aim: Number (%)
Reliability and Validity Data	7 (64%)	7 (58%)
Reliability Data only	2 (18%)	0
Validity Data only	1 (9%)	4 (33%)
No Reliability or Validity Data	1 (9%)	1 (8%)

Tool practicalities

Feasibility

6 out of the 23 articles reported or provided evidence of feasibility. 5 of these reported negative feasibility issues (Table 2-11)

Table2-11: Publications providing feasibility data

Study number	Feasibility
1	Problems with review of videotape in terms of being able to

	see specific actions. Reference to economic and logistic challenges.
12	Requires a research fellow to be present in theatre. Twice the fellow had to scrub up and assist with the operation.
15	Labour-intensiveness of role-play.
16	Discussion regarding reducing the number of elements assessed from 27 to 10 in order to have an acceptable subject burden.
20	Response rate needs to be higher (was 70%).
22	Less costly than high-fidelity simulation

Educational impact

6 out of the 23 articles reported or provided evidence of educational impact. All of these reported positive educational impact. 5 studies had Kirkpatrick Level 1 impact and 1 study had Kirkpatrick Level 2 impact. (Table 2-12)

Table2-12: Publications reporting educational impact

Study number	Kirkpatrick Level	Educational impact
5	1	53 % of students found the comments helpful.
7	1	Majority said it was a valuable learning experience
9	1	Trainee perception of change in skills and knowledge assessed using pre/post testing
15	1	Almost 75% of the students agreed that they had improved communication skills and knowledge.
18	1	The nursing students reported that incorporating the simulation into their class curriculum positively influenced their performance.
22	2	Improvement in performance of participants in a simulated environment.

There is a difference in the number of studies deemed to have provided consequences validity (1) and those providing evidence of educational impact.

The consequences validity has been restricted to the tool only (i.e. did participants provide feedback on the tool), while educational impact has been used to show the effect of the entire intervention (e.g. simulation scenario using tool to assess)

Acceptability

5 out of the 23 articles provided evidence or referred to acceptability of their assessment tool. 4 provided positive responses and one provided a negative response. 3 discussed acceptability in terms of the participants, the other two discussed acceptability in terms of the raters (Table 2-13)

Table 2-13: Publications providing acceptability data

Study number	Participant/ Rater	Acceptability
1	Participant	Feedback from trainees that session and feedback were useful. All would like to participate in similar sessions in the future. Session was realistic.
5	Participant	Comments from focus groups post-intervention suggested that there was a lack of constructive feedback.
7	Participant	Peer assessment valuable and overall assessment was fair.
8	Rater	78-82% found it average to easy to rate using the tool
15	Rater	SHOs who were role-playing the parents and the assessors themselves thought it was acceptable. The SHOs thought this was better than a question and answer approach. (p.198)

Discussion

This review analysed 23 studies for their use of tools which assessed leadership and/or teamwork of individual medical or nursing undergraduates or post-graduates.

Study characteristics

Journals, subject areas and country of study

As might be expected when searching for articles whose subject matter is an assessment tool, medical education journals featured the greatest number of articles. However, it may be surprising to see that surgical journals had published the second greatest number. Crew resource management (CRM) and non-technical skills (NTS) may have jumped the professional divide from aviation into anaesthesia (Gaba et al., 2001) but it seems that surgery has seen more development in this area. Because the retrieved studies looked specifically at the teamwork and leadership of an individual, it may be that studies that looked at the teamwork and leadership of teams may have shown a different subject matter publication profile.

The publication profile by country showed that the UK, with 1/6th of the population of the USA, performed almost as many studies. This may reflect the high quality of research carried out in the UK, as well as the fact that a number of studies used NOTECHS as a template for their assessment tool. NOTECHS was a behavioural marker system developed under the auspices of the European Joint Aviation Authority by, amongst others, Rhona Flin from the University of Aberdeen (Flin et al., 2003). Dr Flin went on to aid in the development of the anaesthesia non-technical skills (ANTS) taxonomy (SN: 8) and the non-technical skills for surgeons (NOTSS) taxonomy (SN: 23). NOTECHS also informed the development of the tools found in study numbers 12, 13, 14, and 19.

Samples

The majority of studies were performed on uniprofessional (medical), postgraduate subjects. It is likely that uniprofessional studies are logistically easier to organise. It is also possible that the search strategy of this review, by

limiting the search to tools which assess individuals, underlined the difficulty of developing a tool which could be used across professional boundaries.

It is unclear why there are no uniprofessional (nursing) studies. A number of papers which were rejected from analysis explore leadership in nursing (Lemire, 2002, Pollard et al., 2005, Bensfield et al., 2008) but none discussed the use of an assessment tool. This may reflect a basic difference in assessments between medicine and nursing training programmes.

Of the 8 studies that assessed undergraduate subjects, two (SN: 15, 18) had both nursing and medical subjects. Both of these studies were short-term studies. The remaining 6 studies were divided between longer-term questionnaire-based studies (SN: 5, 7, 17, 20) and shorter-term observer-based tools (SN: 3, 21).

The most common form of recruitment was convenience sampling. This form of non-probability sampling is not scientifically rigorous but perhaps understandable given the logistic difficulties faced by researchers. In addition, as many of these were exploratory, proof-of-concept studies, the authors may have felt that additional studies could be carried out in the future. The next most common form of recruitment was purposeful sampling which is a more defensible method of recruitment and may allow for greater generalisation of study findings. Somewhat surprisingly, five of the 23 studies did not specify their recruitment method. One study (SN:14) states "There were 20 surgeons who were divided into 2 groups" (p.140), while another (SN:21) states "35 first-year medical students were recruited." (p.31).

The sample sizes varied between 6 and 134. The larger sample-size studies tended to have questionnaire-based assessments, while the smaller sample-sized studies tended to be observer-based assessments. These sample sizes reflect the practicalities of carrying out real-time, observer-based assessments on large numbers of subjects. The two studies which used scripted samples (SN: 8, 23) aimed to remove one of the variables from assessment tool evaluation by

standardising the subjects. In these two studies the subjects "acted out" a script which was then assessed by raters using the assessment tool. The benefits of such a study design are clear, however it does mean that the assessment tool is being used *in vitro*, with a concomitant uncertainty of *in vivo* performance.

Aims and objectives

Just under half of the studies had the development and/or evaluation of an assessment tool as their primary aim. It might be expected that these studies would be more likely to report psychometric properties of the assessment tool, such as reliability and validity. As Table 2-8 (above) shows this is indeed the case. Encouragingly, 11 out of the 12 studies which did not have the development and/or evaluation of an assessment tool as a primary aim still reported some psychometric data.

Study designs

The variety of study designs reflects the different approaches to assessing teamwork and leadership. The pre-/post-intervention studies (SN: 3, 9, 22) are methodologically sound in that they aim to show a change in teamwork and/or leadership due to the intervention with a concurrent demonstration of validity if the assessment tool can show this change. The surveys were generally larger studies (smallest sample size: 95) looking at teamwork and/or leadership over a longer timeframe (Appendix 2-5). The observational study (SN: 4) looked at the leadership of medical senior house officers (SHOs) during real cardiac arrests. The assessment tools used complemented the study design, for example the surveys had a larger number of items for scoring given the completion time available.

The time-frame to which the assessment referred to also varied, from single scenarios to many months. The relevance of this is two-fold. Firstly, the assessment tool we sought was to be used over a short time-frame within a simulated scenario. Secondly, there is an argument that the leadership and/or teamwork displayed and required over longer time-frames is different from that seen in shorter (crisis) scenarios.

The majority of studies used a simulated environment. The benefits of using a simulated environment include standardisation and no risk of harming a patient (Issenberg et al., 2005). One of the drawbacks of simulation is that the environment can, by definition, never be as real as "real-life" and so the assessment of teamwork or leadership performance in the simulator may have reduced validity.

A minority of the studies used peer assessors. As Norcini (2003) states: "(Peer) assessment can be good or bad depending on how it is carried out" (p.539). For our assessment tool, the use of peer assessors in an unfamiliar high-fidelity simulation environment and the possible subjectivity of peer assessment would create another variable requiring compensation.

Assessment tool development and tool types

Tool development

The majority of studies used existing tools, in many cases with modifications. Although the use of an existing tool simplifies the methodology, modifications mean that those tools do not have the same validity or reliability as the original. For example Sevdalis et al. (SN 19) modified a NOTECHS rating system from its original 5-point Likert to a 6-point Likert scale. In their paper they state: "Existing empirical evidence suggests that the NOTECHS rating system can be used reliably in the context of CRM (Crew Resource Management)." However modification of the tool mean that previous data supporting the tool is no longer reliable.

The next most common type of tool development was "author preference", for example Kaye and Mancini (SN 10). In this study the authors wrote: "the team leader must be able to perform in at least five areas: assessment of both status and team performance, dysrhythmia recognition, defibrillation, drug therapy, and trouble-shooting." This is a list of skills and behaviours expected of the leader, with the use of the term "at least" suggesting that there may be more which the authors are not assessing. In addition, using a catch-all phrase such as

"trouble-shooting" along-side very specific skills such as "dysrhythmia recognition" suggests that insufficient thought has gone into tool development. "Author preference" therefore may be considered the least robust tool development method.

A number of other studies, in contrast, used very robust methods to develop their tools. Fletcher et al (SN: 9) carried out a literature review and cognitive task analysis interviews, which resulted in a prototype taxonomy. This taxonomy was amended using anaesthetic incident reports, observations in theatre and results from an attitude survey of anaesthetists. Yule et al. (SN: 23) carried out a literature review, observations in theatre and cognitive task analysis interviews with experts. They also examined surgical mortality reports and undertook an attitude survey of theatre staff. Both of these studies required significant input in terms of time and money, but the resultant assessment tools are more evidence-based than those where the authors themselves decided what to assess.

Tool types

The majority of studies used a Likert scale. The number of points on the scale varied from 4 to 9. Two factors are relevant here. The first is that as the number of points and therefore possible responses increase, the poorer the test-retest and inter-rater reliability (Preston and Colman, 2000). The study using the 9-point Likert scale (SN: 17) did not report reliability data. The second is that Likert scales with odd numbers of points allow for a middle value (e.g. acceptable, neutral, neither good nor bad), while even-numbered scales force assessors to decide whether or not a behaviour was on the good or the bad side of the spectrum.

The number of Likert items is influenced by the amount of time available for scoring. A questionnaire survey (SN: 16) study may have 27 Likert items, while a scenario-based assessment (SN: 20) may only have 5. The number of Likert points therefore must be sufficient to provide a detailed assessment of a given behaviour, the choice between even and odd item numbers must be an informed

one and the number of Likert items will primarily depend on the time available to the assessor.

A minority of studies used a checklist. A checklist in its simplest form might use a "performed/not performed" assessment method. The benefits of this seem clear, eliminating the need of the rater to provide a subjective assessment, e.g. they performed well or very poorly. However, checklists harbour a number of pitfalls. For example, if an action, such as checking the blood pressure, is performed only once in a scenario but actually should have been checked a number of times then the question "Blood pressure checked" becomes difficult to answer. The rephrasing of the question to "Blood pressure checked when appropriate" eliminates this problem but re-introduces the subjectivity that checklists are meant to remove. This attempt to have the best of both worlds is evident in SN 9 which uses a checklist where three possibilities are defined: 1) Performed 2) Not performed and 3) Borderline. A study by Regehr et al (1998) supports the use of a global rating (Likert) scale over checklists, showing better reliability and validity in the hands of expert raters. In addition, checklists may fail to differentiate between expert and novice, as experts use recognition-primed decision-making which relies less on a checklist-type approach to problem-solving (Flin, 1996).

Tool psychometrics

Validity

The majority of tools provided some validity data, the exceptions were SN: 15, 18, and 23. Two tools (SN: 3, 9) claimed content validity by the fact that they were based on existing validated tools, although (as discussed above) this is not necessarily the case. Two studies (SN: 1, 11) provided more robust content validity by being examined by content experts.

Overall the validity evidence provided was uni-dimensional and therefore insufficient to recommend the use of any one assessment tool by validity evidence alone.

Reliability

Only 16 studies provided reliability data and there was great variability both in approach to ensuring reliability (via rater calibration) and presenting the evidence. With regards to the former, some studies had no reference to rater calibration (SN: 2, 5, 11, 19, 21), referred to rater calibration but did not specify what this involved (SN: 12) or provided minimal rater calibration (e.g. 1 or 2 videos). Other studies provided varying degrees of rater training, from 4 hours (SN: 8) to 5 videos (SN: 13, 14). A study (SN:23) which describes poor inter-rater reliability and a need for more in-depth training and calibration of raters used 3 videos to calibrate. In terms of calibration, it would seem prudent to use at least 5 videos (or other observations) to standardise the raters.

A variety of methods were used to express inter-rater reliability: Intra-class correlation (4), Cronbach's α (2), Cohen's K (2), r_{wg} (2), Pearson correlation coefficient (1), generalizability co-efficient (1) and percentage agreement (1). The statistical literature regarding inter-rater reliability, and the methods with which to assess it, is complex. The selection of one tool over another is often a matter of opinion (Gisev et al., 2013). The different scoring measures are explained below (Downing, 2004, Cook and Beckman, 2006, Gisev et al., 2013)

Table 2-14: Inter-rater reliability scoring systems

Measure (Reference)	Definition	Comment
Intra-class correlation (Shrout and Fleiss, 1979)	Uses analysis of variance (ANOVA) to estimate how well ratings from different raters coincide. ICC = Between subjects variance/(Between subjects variance + Within subjects variance)	Can be used to calculate the actual reliability of the n-raters as well as the reliability of a single rater. Compensates for missing values.
Cronbach's α	The expected correlation of two	Normally used for test-

(Cronbach, 1951)	tests that measure the same construct. An internal consisteny coefficient.	retest data, but can be used for a single test given the assumption that the test is measuring a single construct (and single test is split into 2 halves)
Cohen's K (Cohen, 1960)	Agreement corrected for chance K = (Proportion observed agreement – Proportion expected chance agreemen)/(1 – Proportion expected chance agreement)	Has been used as measure of both inter-rater agreement and inter-rater reliability. Should only be used for binary data. Values from -1 to +1.
r_{wg} (James et al., 1984)	Inter-rater agreement	Does not account for agreement which occurs by chance
Pearson correlation coefficient (Pearson, 1895)	Calculates correlation rather than agreement	Should be used for test-retest or alternate forms reliability. (Can have perfect correlation (all points on one line) without agreement (all points on line of equality y=x)(Gisev et al., 2013)
Generalisability co-efficient (Cronbach et al., 1963)	A random effects theory which aims to identify all sources of variation.	The most "elegant" method of assessing inter-rater reliability (Gisev et al., 2013)
Percentage agreement	Percent of identical responses % = (Number of concordant	Does not account for agreement which occurs by chance. (Gisev et al.,

	responses/Total number of responses) x 100	2013)

The above table suggests that the choice of the Pearson correlation coefficient by SN 21 was a poor one. In addition, those studies which used r_{wg} (SN: 8, 23) or percentage agreement (SN: 1) were assessing inter-rater agreement rather than inter-rater reliability. They do not account for the effects of chance, nor do they take into account the proximity or disparity of a score along a Likert scale. For example if rater 1 scores a 5 and rater 2 scores a 4 this would provide the same lack of agreement as a score of 5 and a score of 0. However, the former scoring pattern would be taken into account by an inter-rater reliability measure.

Gisev et al. (2013) provide a useful table (Figure 2-3), adapted from Tinsley and Weiss (1975), which shows how inter-rater agreement and inter-rater reliability will provide different results for different data. For example, one can have high inter-rater agreement but poor inter-rater reliability, and vice-versa.

Figure 2-3: Inter-rater reliability (IRR) and agreement (IRA) (p.332, Gisev et al., 2013) [1]

Hypothetical ratings of communication skills (on a scale of 1 to 10) of 10 pharmacists illustrating different levels of IRR and IRA for interval-scaled data

Pharmacist	Case 1 (high IRR and high IRA)			Case 2 (high IRR and low IRA)			Case 3 (low IRR and high IRA)			Case 4 (low IRR and low IRA)		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
A	1	1	1	1	3	6	5	6	5	1	5	10
B	2	2	2	1	3	6	4	4	4	1	6	9
C	3	3	3	2	4	7	6	6	6	2	5	10
D	4	4	4	2	4	7	4	5	6	2	6	1
E	5	5	5	3	5	8	5	4	4	3	6	5
F	6	6	6	3	5	8	6	6	5	7	4	9
G	7	7	7	4	6	9	4	4	5	3	5	8
H	8	8	8	4	6	9	5	5	4	3	5	2
I	9	9	9	5	7	10	4	5	3	1	6	7
J	10	10	10	5	7	10	6	6	6	6	3	9

Adapted from Tinsley and Weiss.

One of the studies that used Cohen's kappa (SN: 1) did so appropriately, as the study used binary data, however the other study (SN: 4) used Cohen's kappa for ordinal data when they should have used a weighted kappa instead. A weighted

[1] Reprinted from Research in Social and Administrative Pharmacy, 9/3, Natasa Gisev,J. Simon Bell,Timothy F. Chen, Interrater agreement and interrater reliability: Key concepts, approaches, and applications, 330-338., Copyright (2013), with permission from Elsevier

kappa needs to be used for ordinal data as it takes into account the variation in distance between interval points (i.e. the distance between a score of 1 (very poor) and a score of 2 (poor) is not the same as the distance between a score of 2 (poor) and a score of 3 (average)) (Gisev et al., 2013)

Gisev et al. (2013) also provide a useful table showing the measures they consider appropriate for a given dataset and number of raters.

Figure 2-4: Interrater indices, level of measurement and number of raters (Gisev et al., 2013)[2]

Examples of interrater indices suitable for use for various types of data[a]

	Level of measurement					
	Nominal/categorical		Ordinal		Interval and ratio	
	2 raters	>2 raters	2 raters	>2 raters	2 raters	>2 raters
Interrater indices	Cohen's kappa	Fleiss' kappa	Weighted kappa	Kendall coefficient of concordance	Bland-Altman plots	ICC
	ICC Weighted kappa	ICC	ICC	ICC	ICC	

[a] Table is not exhaustive and represents a summary of some of the indices and the contexts in which they can be used only.

The majority of studies lacked detail when presenting inter-rater reliability scores. Some studies (SN: 22) provide a single intra-class correlation (ICC) score for the entire dataset without specifying which particular data the score refers to, while others provide a reliability score for a single item (such as a total score or global rating score) only, without detailing the reliability data for the remainder of the Likert items (SN: 1, 12, 13).

Table 2-15 below illustrates the relationship between a given score and accepted reliability .

Table 2-15: Inter-rater indices and acceptable scores

Measure	Comment
Intra-class correlation	>0.9 for high-stakes, 0.8-0.9 for moderate stakes, 0.7-0.8 for low-stakes (Downing, 2004)
Cronbach's α	>0.7 is adequate, although lower values have sometimes been

[2] Reprinted from Research in Social and Administrative Pharmacy, 9/3, Natasa Gisev,J. Simon Bell,Timothy F. Chen, Interrater agreement and interrater reliability: Key concepts, approaches, and applications, 330-338., Copyright (2013), with permission from Elsevier

	considered acceptable (Nunnaly cited by Sevdalis 2008)
Cohen's K	<0.00 poor, 0.00-0.20 is slight, 0.21-0.4 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), 0.81-1.00 (almost perfect) (Landis and Koch as cited by Gisev2013)
r_{wg}	>0.7-0.8 is acceptable (Nunnaly cited by Yule2008)
Pearson correlation coefficient	0.4-0.59 moderate; 0.6 slightly higher (Wright2009)
Generalisability co-efficient	>0.9 for high-stakes/summative, 0.8-0.89 for moderate stakes, 0.7-0.79 for low-stakes/formative (Downing, 2004)
Percentage agreement	No data

8 out of the 12 studies that provided inter-rater reliability data had acceptable (>0.7) reliability for formative assessments and 8 out of 9 studies that provided internal consistency data had acceptable (>0.7) internal consistency. However, the selective publication of data, the use of inappropriate statistical tests and the paucity of reliability data within the 23 studies provides a challenge to the selection of a tool for assessment of leadership and teamwork in medical undergraduates.

Tool practicalities

Feasibility

The lack of reference to feasibility in the majority of the studies makes analysis difficult. The six studies which did refer to feasibility issues did so indirectly and generally only in a single sentence. The five studies referred to different categories of problems:

- Equipment (SN 1)
- Staffing (SN 12: data collection, SN 15: creation of testing environment, SN16: test burden)
- Response rate (SN 20: insufficient respondents)
- Financial (SN 1)

These issues are perhaps not surprising and must be considered in the development of any assessment tool. The lack of reference to feasibility issues in the other studies is disappointing. It is likely that either there were issues regarding feasibility or that there would be issues if the assessment were carried out on a larger scale or repeatedly, but the studies do not elaborate.

Educational impact

Only 6 studies discuss the educational impact of the study and only 1 of these achieves Kirkpatrick Level 2 impact. It is disappointing that more of the studies did not provide educational impact information. Although assessing for Kirkpatrick Level 2 impact (and above) requires a larger (and longer) study, assessing for Level 1 impact would normally only entail a post-assessment questionnaire. The failure to collect this data means that an important part of supporting information is lost.

Acceptability

The acceptability of a tool can refer to how acceptable it is to the raters, to the participants or both. Only 5 studies discuss acceptability. 3 of these refer to participant acceptability and 2 refer to rater acceptability. Once again, the majority of studies fail to collect relatively simple and straightforward information which would support the use of a given tool. An assessment tool must be acceptable to raters so that they will be willing to use it and will attempt to rate to the best of their abilities. An assessment tool must be acceptable to participants so that they will believe they are being treated fairly.

Limitations

This systematic review had a number of limitations. The search term restrictions may have resulted in the exclusion of some relevant studies. The initial search resulted in 3014 articles, and time constraints led to an initial exclusion strategy based on title alone. This method has been described elsewhere (Holly and Salmond, 2011, Centre for Cognitive Ageing and Cognitive Epidemiology, 2013) and the process was in favour of keeping a source,

however it may have led to the exclusion of relevant studies (For evaluation, Appendix 2-8 details the first 25 articles rejected by title alone.) In addition, the title and abstract reviews were carried out primarily by the MD student, and this may have resulted in the loss of relevant studies. The assessment of teamwork and leadership was a relatively new concept and it is possible that some studies assessed elements of teamwork or leadership without specifying either of the two terms. These studies would not have been included in our search strategy.

Conclusion

This literature review made use of a robust search strategy and encompassed the databases where relevant articles might be reasonably expected to be found. The 23 studies which matched the inclusion and exclusion criteria were a diverse assortment. Although all assessed some component of teamwork or leadership, the study design and data quality varied greatly. The reporting of tool psychometrics was, on the whole, poor. The variation in the published reliability data may, in part, be due to the fact that there is greater variability between raters when observing non-technical skills (such as teamwork and leadership) than technical skills (such as cannulation or suturing)(Yee et al., 2005). It is also possible that the authors chose the inter-rater reliability test that provided the "best" score. By the same token it is possible that the rationale for not displaying the entire inter-rater reliability for every observation but rather choosing a global rating score, is because the inter-rater reliability for such global scores is better than for individual scores.

The reporting of tool practicalities such as feasibility and cost-effectiveness was worse than the psychometric data. It is unclear why more studies did not include at least some discussion around how their tool would be used *in vivo*.

Our literature review was designed to unearth tools which were used to assess teamwork and/or leadership in healthcare professionals. Our own study into the assessment of teamwork and leadership of medical undergraduates was going to be simulation-based. This means the tool would have to be concise enough to be completed during a 15-20 minute scenario; tools which were looking at longer time-frames would probably be less relevant. Of the 23 studies, 16 fit the criteria of looking at a short interaction.

In addition, we would be focusing on medical undergraduates, so tools assessing this group might be considered more relevant. Of the above 16 studies, 5 match the criteria of being short-term and undergraduate (SN: 3, 15, 18, 21, 22). If we were to specify that our raters were to be non-peers and the interventions had to be simulated, then the same 5 studies would still match these criteria. Of

these 5 studies, SN 18 is the only study which aimed to assess both teamwork and leadership. Unfortunately none of the 5 studies matched all our requirements for a usable assessment tool. SN 3 claimed "face validity" without providing any evidence, while SN 15 and 18 provided no validity evidence at all. SN 21 and 22 provided some construct/relationship to other variables validity but nothing more. SN 3, 15 and 18 provided no reliability evidence. SN 21 provided inter-rater reliability evidence using a poor statistical test (Pearson correlation). SN 22 provided inter-rater and internal consistency data, but the inter-rater data was just within the acceptable range, an ICC of 0.71.

The systematic review provided us with the following fundamental considerations:

- "Author preference" is not an acceptable method of tool development
- Validity and reliability evidence has to be considered from the outset
- Tool practicalities such as feasibility, acceptability and cost-effectiveness have to be investigated and reported

The systematic review made it clear that there was no "off the shelf" assessment tool which could be used in our study. In practice this meant that we would have to use the literature review to inform the development of our own assessment tool. This process is expanded in Chapter 4: on "Development and Evaluation of the Assessment Tool".

Chapter 3: Focus group study

Introduction

What are medical students' views on professionalism, teamwork and leadership? Stakeholders, such as the General Medical Council (GMC) and the Medical Schools Council, set standards and supply definitions; however there has been little research into the students' own beliefs and attitudes. This chapter will provide an overview of the literature and its limitations, followed by a rationale for the use of focus groups. The body of the chapter will present the methods, results and discussion. The chapter's conclusion will place this work into the context of the existing literature.

Existing research and its limitations

The papers referred to below are not meant to be an exhaustive list of all research that has taken place within the parameters of professionalism and medical undergraduates. They are instead a selection of representative studies which will inform our understanding of the current paradigms.

Rennie and Crosby (2002) examined the attitudes of medical undergraduates in a Scottish medical school to whistle-blowing in the context of academic misconduct. They found that a minority of medical students would whistle-blow and, as medical students progressed, they were less likely to do so. Rennie and Crosby argue that students have a duty to whistle-blow, as a precursor to the self-regulation expected of the medical profession. They found that students feared retaliation, acted in self-preservation, and that there was an increased practice and acceptance of misconduct as they progressed. In relation to our study, the only limitation of Rennie and Crosby's work is that they focused on a very specific aspect of professionalism.

Brainard and Brislen's (2007) paper is not so much an apologia as it is a defence of medical students and an attack on the "hierarchy of academic authority" (p.1010) in the United States. They assert that the short-comings in medical education, the subjective assessment of professionalism and the lapses in professionalism they witness, result in confusion. Brainard and Brislen call for

the hidden curriculum to be addressed, professional instruction and objective evaluation of professionalism. Unfortunately Brainard and Brislen's paper is a narrative, quasi-editorial. They do not provide in-depth data on their techniques and the abstract states: "Their views on professionalism education, although not the result of qualitative research…." (p.1010). Although one cannot totally refute their paper as it is a view of professionalism based on their experiences, the lack of scientific rigour makes their conclusions somewhat unsupported.

Chard et al.'s (2006) study was based on a questionnaire which was sent out to medical trainees and medical students in the UK. There is no mention of the hidden curriculum or unprofessional behaviour by peers or seniors. Instead the authors conclude that the main threats to professionalism were the "expectations of the public and politicians set in the context of limited financial resources, changes in working patterns, protocol-driven care, and changes in medical education" (p.69). They also state that the respondents felt medicine was a profession which was defined by responsibility to patients. Respondents also thought that the "standards of care should be defined and regulated by the profession, and that training should be directed by the profession" (p.69). There are two main methodological limitations to this study. The first is that the responses were not separated into under- and post-graduate. 20% of the respondents were medical students but it is not known how or if their responses differed from post-graduates. The second limitation is that, as with all questionnaire surveys, the responses are limited to the questions being asked. For example, 97% strongly agree or agree that medicine is a profession, but the statement "Medicine is a job" is not posed. One cannot know what the response to this statement would have been, nor if everybody understood what was meant by the word "profession".

Feudtner et al. (1994) sent a questionnaire to 3rd and 4th year medical students in Pennsylvania, USA. The responses concur with the work of Rennie and Crosby (2002) and Brainard and Brislen (2007). Medical students carried out and witnessed unprofessional behaviour, and the main reasons for being unprofessional was because of fear of poor evaluation or to fit in with the team.

Feudtner et al. also comment on the harm that exposure to unethical behaviour causes in terms of the erosion of ethical principles. This attrition has been detailed by others such as Herbert et al. (1992).

Ginsburg et al. (2003) conducted interviews with 4th-year medical students. Based on analysis of the transcripts the authors found that students were motivated according to principles, affects or implications. Implications of behaviour were the dominant motivating factor, in particular "disavowed" implications such as concern about grades and assessments. Ginsburg et al. suggested that this "disavowed curriculum" needed to be studied, understood and dealt with. A drawback of this study is that the authors looked at a specific element of professionalism, the imagined response to unprofessional behaviour in video-taped scenarios, rather than professionalism as a whole.

Hicks et al. (2001) analysed focus groups with 4th-year medical students and categorised the ethical dilemmas that students are faced with. They specify 3 types: 1) Conflict between medical education and patient care, 2) Responsibility exceeding student's capacity and 3) Involvement in care perceived to be substandard. The authors do not provide any data on the students views on professionalism. In addition, because the paper is only a page and a half long, the methodology section merely states that they carried out a content analysis.

Jha et al. (2006) conducted interviews with a range of people involved in medical education, including doctors, allied health professionals and medical students. After thematic content analysis, the authors hypothesised two types of professionalism: conceptual and behavioural. In addition they found that seven themes arose from the data: compliance to values, patient access, doctor–patient relationship, demeanour, professional management, personal awareness and motivation.

The title of van Rooyen's (2004) paper "The views of Medical Students on professionalism in South Africa" sounded promising. However the paper itself refers to the views of students on a charter published in the Annals of Internal

Medicine (The ABIM Foundation et al., 2002) and whether they feel the charter should/could be applied to South Africa. There is no exploration of the students' own thoughts on the matter of professionalism.

Lastly, Leo and Eagen (2008) refer to professionalism surveys and focus groups by a number of US medical schools, but unfortunately do not provide any references to these studies. Emails to the authors did not receive a reply.

Current paradigms

The quality, rigour and generalizability of medical education research have been subjects of criticism (Carline, 2004) and medical education research has been seen as the "poor relation" to medical research (Todres et al., 2007). To some extent this state of affairs has been due to the tension between those who see medical education research as a social science and those who see it as firmly placed within the biomedical science setting (Bunniss and Kelly, 2010). The former may prefer qualitative methodologies, while the latter may focus on the quantitative.

Medical education therefore has researchers approaching the subject from a wide range of philosophical stances. According to Bunniss and Kelly (2010), current paradigms include positivism, post-positivism, interpretivism and critical theory (Table 3-1)

Table 3-1: Current paradigms in medical education

	Positivism	Post-positivism	Interpretivism	Critical theory
Ontology (The nature of reality)	Reality is objective. There is one truth	Reality is objective. There is one (most probable) truth.	Reality is subjective. There is no one truth.	Reality may be objective but truth is contested.
Epistemology (The nature of knowledge)	Knowledge is neutral, value-free and objective (e.g. Objectivism)	Knowledge may not be fully accessible.	Knowledge is subjective. There is no "correct" way of knowing. (e.g. Subjectivism)	Knowledge is co-created and constantly revised (e.g. Constructionism)
Methodology (The nature	Predict and control.	Falsify hypotheses, concepts and	Inductive, diverse interpretations.	Emancipatory, diverse and

of the approach to research)	Deductive.	variables well-defined.		under-represented views.
Methods (Data gathering techniques)	Quantitative (RCTs, questionnaires)	Quantitative and Qualitative (Surveys, interviews, focus groups)	Qualitative (Observation, interviews)	Quantitative and Qualitative (Focus groups, observations)

In his book "The Foundations of Social Research" Crotty (1998) advocates a slightly different framework, consisting of four elements:

1. Epistemology: the theory of knowledge embedded in the theoretical perspective and methodology
2. Theoretical perspective: the philosophical stance informing the methodology
3. Methodology: the strategy behind the use of particular methods
4. Methods: the technique used to gather data

Crotty acknowledges the fact that there is a degree of confusion in the research community caused by the lack of consistency in the terminology. Ultimately, whether one considers Interpretivism to be a paradigm (Bunniss and Kelly, 2010) or a theoretical perspective (Crotty, 1998), the critical task is to ensure that the philosophical approach to the research has been clearly defined.

With that in mind, the philosophical approach adopted in this MD was "critical theory". Critical theory aims to explore the "construction of knowledge and the organisation of power… in institutions such as schools, hospitals and governments…" (p.633) (Reeves et al., 2008) The power structure within Medicine means that the views of certain groups, such as medical students, are under-represented. The focus groups provided an opportunity for their views to be heard. Using medical students' views on teamwork and leadership allowed us to develop an assessment tool which they would find acceptable. The "think

aloud" allowed us to gain insight into the reasoning behind their unwillingness to challenge authority.

Figure 3-2: A definition of constructionism (Crotty, 1998) (p.42)

Constructionism:
all knowledge, and therefore all meaningful reality as such, is contingent upon human practices, being constructed in and out of interaction between human beings and their world , and developed and transmitted within an essentially social context

The epistemology of constructionism argues that truth and meaning are constructed by our engagement with the world (see Figure 3-1). Constructionism rejects both subjectivism and objectivism, truth and meaning are instead formed from the interaction between subject and object. In this sense, teamwork and leadership are social constructs. The research methodology relies on dialogues, between the researcher and others, between research participants or, as in the "think aloud" technique, within the same research participant. In this chapter, the use of focus groups allowed for the exploration of the thoughts and beliefs of medical students. In addition, although generalisability is discussed, the existence of multiple realities, as hypothesised by constructionism, means that the themes are not considered to be in any sense "universal". In Chapter 4, the assessment tool is constructed from a number of sources, and the evaluation of the tool, although providing psychometric data, lends equal weight to its acceptability and feasibility. The use of the "think-aloud" technique in Chapter 5 again allows for the construction of meaning and truth by the participants when considering their own behaviours and actions.

Rationale and use of focus groups

Focus groups developed out of work by Emory Bogardus in 1926 and later work by Robert Merton and co-workers who wanted to examine the impact of persuasive messages in World War II (Frey and Fontana, 1993, Kitzinger, 1994, Asbury, 1995). Focus groups were primarily used for consumer analysis and product evaluation and did not feature in medical education research until the 1980s (Stalmeijer et al., 2014). Although they have been used for a variety of purposes, the basic characteristics of a focus group are "a semi-structured group session, moderated by a group leader, held in an informal setting, with the purpose of collecting information on a designated topic" (p.413) (Carey, 1995b).

This designated topic is therefore the "focus" of the group (Powell and Single, 1996).

The use of focus groups is strongly supported within a constructionist research approach. In contrast to, for example, a questionnaire, a focus group allows people to reflect on their responses (Dolan et al., 1999), and to "describe the rich details of complex experiences and the reasoning behind their actions, beliefs, perceptions and attitudes" (p.124) (Carey, 1994). The benefit of a focus group over an interview is that the social dynamic can allow the exploration of subjects which would not arise during a one-to-one interaction between researcher and interviewee (Carey, 1994). Focus groups are ideal for generating data on group norms (Bloor et al., 2001) or to quote Kitzinger (1995), focus groups tell you "not only what people think, but how they think and why they think that way" (p. 299) and "reach the parts that other methods cannot reach" (p.299). Of course, one cannot assume that every (or any) focus group will reveal what people "really" think but instead one must analyse the data with an understanding that this is a "public discourse about a topic" (Smithson, 2000) (p.114) which may reveal underlying beliefs and attitudes.

According to Morgan (1988), focus groups can be used to "explore new research ideas or to examine well-known research questions" (p.24) or "as preliminary research to prepare for specific issues in a larger project" (p.24). Both of these possibilities were of utility to this project. As detailed above, there had been a number of studies regarding professionalism and medical students, but very few on how medical students themselves felt about the global concept of professionalism. In addition the "larger project" of assessing teamwork and leadership in medical undergraduates would benefit from an exploration by them into the meanings of the terms teamwork and leadership.

Some researchers, such as Nicolson and Anderson (2003), do not agree with running focus groups to inform the development of a tool. They argue that the focus group material loses its depth and richness by becoming simplified elements within a tool. However, Sim and Snell (1996) and Carey (1994) claim

that focus groups may be particularly helpful in the development or refining of tools or instruments. This is also Thomas et al.'s (2004) rationale for using focus groups in the development of a behavioural marker system. Lastly, Barbour (2005) asserts that tool development as one endpoint does not preclude in-depth analysis of the focus group discussions.

In summary, it was decided to use focus groups because they would provide us with insights into the views of medical undergraduates with respect to professionalism, teamwork and leadership. The focus groups could serve a dual nature of being an exploratory method (for tool development) and being useful for eliciting the student perspective, particularly with reference to the "hidden curriculum" (Barbour, 2005). In addition, Wear and Kuczewski (2004) call for a dialogue with medical students and certainly ascertaining these views and exploring any differences between the students and stakeholders might help in the adoption and adaptation of standards. Also, gaining insight into medical students' mental constructs of teamwork and leadership may help in the development of a tool to assess these two skills. Lastly, Morgan and Krueger (1993) recommend the use of focus groups as "a friendly research method that is respectful and not condescending to your target audience" (p.18).

Methods

In their review of focus group research, Twohig and Putnam (2002) quote Hoddinott and Pill:

> "The relationship between the subject and the interviewer, together with the context in which the interviews take place, are important details in appraising qualitative research. A published paper should provide sufficient methodological detail for a reader to be able to replicate the study and confirm the findings if required" (p. 279)

The Methods below may at first glance seem to impart too much detail, however the quote above has acted as a reference point when deciding how much information to provide.

Ethics

Ethical approval for the study was sought and obtained from the University of Liverpool's School of Medical Education Research Sub-Group. As discussed by Smith (1995), ethical considerations in focus groups include the possibility of overdisclosure by participants, as well as the possibility of disclosure of unethical, unprofessional or even illegal acts or behaviours.

In terms of overdisclosure, focus group participants were assured that no identifiable data would be shared outwith the research group. They were also cautioned that, although participants had agreed not to disclose information outwith the group, the author could not guarantee this and that they must therefore consider this before sharing information which may be personally damaging. The approach recommended by Smith (1995) is to allow disclosure and discussion of unethical, unprofessional or illegal acts but to ensure that, if the group did not explore the issues sufficiently, the moderator would have had a later, private chat with the person alleging the act. This would allow the moderator to discuss the participant's rights and responsibilities as a medical undergraduate.

Conceptual framework

According to Fern (2001), there are seven components of a conceptual framework (Figure 3-2) in focus group research. Each is addressed in turn below.

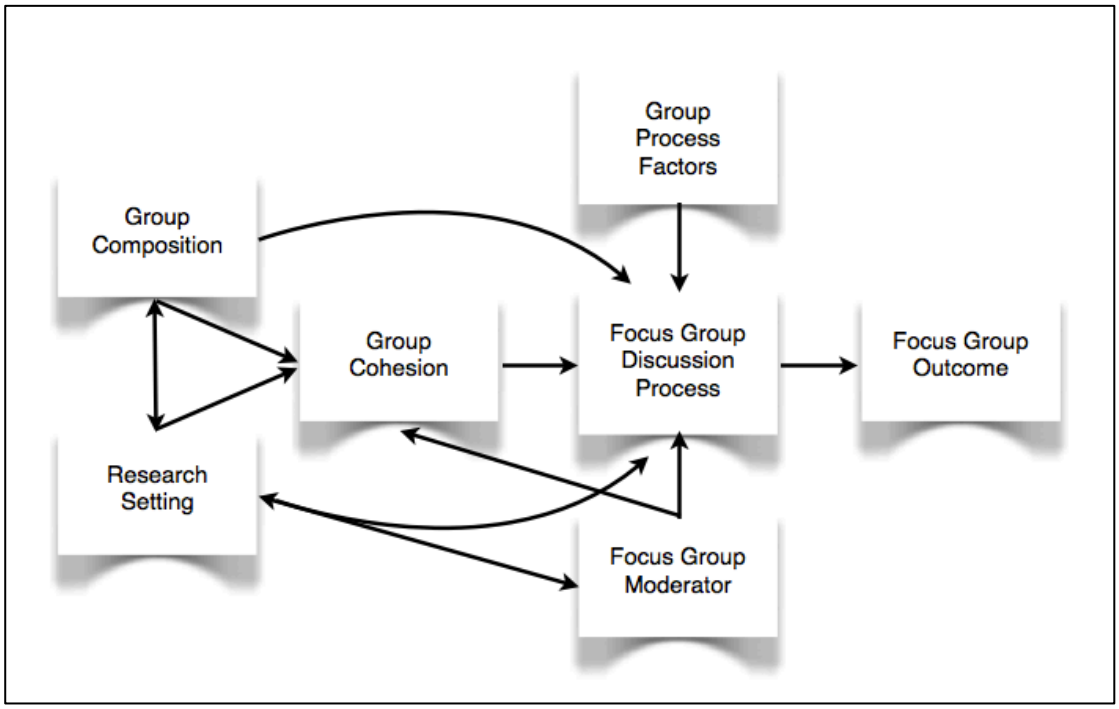


Figure 3-2: Seven components of a conceptual framework (Fern (2001))

1) Group composition

There are six individual characteristics that can describe focus group participants: value orientation, social status, race/ethnicity, age, gender and personality. Group composition therefore depends on the sampling model and sample size.

a) Sampling model

We used a theoretical sampling model, rather than a representative sample (Mays and Pope, 1995). Theoretical sampling is used to answer a particular question or test a particular hypothesis (Kitzinger, 1995). For us the particular area of interest was the assessment of teamwork and leadership in medical students. However, our simulation-based study was going to have to ensure that the medical undergraduates had the prerequisite skills and knowledge to deal

with a simulated clinical emergency and had enough exposure to professionalism, teamwork and leadership to have a fruitful discussion. In effect this limited us to fourth- or fifth-/final-year students. Furthermore, as this was a two-part study, 1) focus groups and literature review then 2) simulation exercise, timing was of crucial importance. If we had recruited final-year medical students in the first part of the study then they would have graduated (and dispersed) before the second part. This meant that we restricted our focus groups to 4th-year medical undergraduates, who became 5th-year medical undergraduates in the second part of the study.

We did not carry out targeted sampling of student sub-groups, e.g. mature students, ethnic minority students, students with disabilities, etc. One of the essentials of focus group research is that the participants have a common experience which they can discuss and which allows them to participate in the group (Asbury, 1995). We decided to focus on this common experience of progressing through four years of medical school, with exposure to a variety of professional behaviours, and styles of teamwork and leadership, rather than differences between the individuals. Barbour (2005) supports this view by calling for enough heterogeneity within a group to provoke discussion but also enough homogeneity to allow comparative analysis between groups to take place.

A website was set up to allow participants to book themselves onto a focus group. The identity of the other participants was not revealed, so as to prevent groups of friends from coming along to the same focus group with a risk of biasing the results. Morgan (1997b) argues that although focus groups consisting of friends may have a more easy-flowing discussion, this will in part be due to the fact that they are relying on "taken-for-granted" assumptions. And one goal of the focus group is to examine these assumptions, which would be difficult if they are not voiced. It has been suggested that the ideal focus groups consist of people with a shared experience who are total strangers to one another (Powell and Single, 1996) but this was not going to be achievable in this population of medical undergraduates. Krueger (1995) meanwhile argues that

successful focus groups can be carried out with people who know one another. We felt that our method of anonymising the booking process and ensuring that all emails were sent to individuals rather than to groups of individuals, provided the correct balance within this population.

Other factors, such as whether or not to offer single-sex focus groups (Morgan, 1997b) were also considered, but it was felt that the discussion and the participants' comfort would not be affected by having mixed-sex groups. In addition, Goldman and McDonald (quoted in (Twohig and Putnam, 2002) challenge the automaticity of segregating according to sex:

> "Traditionally men and women have been segregated in group interviews on [various] assumptions . . . All of these concerns may have once been valid but social observation and actual research experience indicate that these issues are far less relevant now than they were 20 years ago. Today, it is largely inertia and research ritual which perpetuate an automatic segregation of the sexes in almost all group interview projects" (p.281)

The website was set up to allow a maximum of 8 people to attend any given focus group after which participants would be asked to choose a different date. The focus groups dates and number of attendees are provided in the Results section.

b) Sample size

The ideal number of focus groups is debatable. One should continue to run focus groups until no new ideas, beliefs or attitudes are expressed; that is, "data saturation" has been reached (Basch, 1987, Krueger and Casey, 1994). The literature suggests that 3 or 4 (Asbury, 1995, Krueger, 1997, Barbour, 2007) groups are normally sufficient to reach this point. In addition, focus groups which are relatively structured in terms of the questioning (see Focus group operation below), tend to require fewer groups in order to reach saturation (Morgan, 1997b). As a precautionary measure, it was decided therefore to run 5 focus groups.

According to Kitzinger (1995) the ideal focus group size is between 4 and 8 participants. Barbour (2007) suggests a maximum of 8 people per focus group for health sciences research, while Krueger (1995) and Bloor et al. (2001) state that the most effective focus groups consist of 6 to 8 people. Taking an average of 7 people per group, this would mean recruiting 35 participants.

There were 338 fourth-year students at the University of Liverpool Medical School in January 2009. For logistical reasons, i.e. practicalities of travelling, we excluded the 67 students who were based at the satellite campus in Lancaster. For reasons of confidentiality the medical school administration was unwilling to provide us with a list of all 271 students based in Liverpool. We therefore requested to be provided with 60 names and email addresses. These were selected by the year group administrator by taking every 5th name on the year group list resulting in 54 email addresses. The administrator then selected every 4th name for another 6 students to provide us with a total of 60. It was thought that this was a sufficiently random process of selection from within the sample. Although systematic random sampling is not essential for focus group operation, as the goal is not generalizability (Bloor et al., 2001, Lingard and Kennedy, 2007), students are often grouped together by surname and we wanted to ensure that we did not inadvertently select a number of pre-existing groups.

Every one of these students was sent up to 3 emails to invite them to attend one of the focus groups (Appendix 3-6 provides the template for the first email). It has been suggested that an incentive can help with recruitment and retention (Morgan, 1995, Beyea and Nicoll, 2000b) and therefore an incentive of being placed in a draw for a personal mp3 player (RRP: £59.99) was offered. Discussions with a pilot group of students, and researchers at CEDP, suggested that this incentive was sufficient to encourage participation without being disproportionately valuable.

Students were not sent the 3 emails if they either accepted or declined to participate in the study before the next round of emails were sent. The number recruited was less than that calculated to be necessary (see Results) and we asked the year administrator to send us email addresses of an additional 60 students. The randomisation process was as before, with every 4th email address provided and then every third until 60 names were reached. One of these email addresses was a duplicate, so 59 invites were sent out with the same method as before.

The students who had agreed to attend were sent another email containing a participant information sheet (Appendix 3-2) and consent form (Appendix 3-3). They were asked to read the two documents and reply with any queries.

2) Research setting

a) Place

A conscious decision was made to host the focus groups at the Liverpool Medical Institution, a building that would be known to the students but would not be associated with the medical school. Feedback from the 4th year undergraduates who were involved in the questionnaire design had suggested that the students might talk more freely outside University premises. In addition, we felt that the LMI struck the right balance in terms of backdrop and ambience. It is an august, learned establishment, which might naturally induce reflection on medical professionalism and, at the same time, provides a relaxed and welcoming atmosphere. The LMI is also central in Liverpool and a short distance away on foot from both the medical school and the Royal Liverpool University Hospital.

The focus groups were held in the President's Room. The chairs were arranged in a circle, as recommended by Beyea and Nicoll (2000c), around a table, as recommended by Tiberius (2006) with food and drink (tea, coffee, water, juice) provided in the middle of the table. These refreshments were provided in order to ensure the physiological needs of the participants were met (Carey, 1994), especially as the focus groups took place over lunchtime or in the evening (see

Time below). Catering was booked for half-an-hour before the start of the focus groups so that participants could have a bite to eat and pour their drinks. This meant that the focus groups themselves were not interrupted by participants serving food or drink but also allowed for an "icebreaker" period (Powell and Single, 1996) during which informal interaction could occur. The food and drink was placed in the middle of the table so that those participants who did not arrive early could still avail themselves of the refreshments while remaining seated within the focus group circle. In addition, the food was bite-size and not crunchy, the former meant that each item is quick to swallow and the latter reduced the risk of sound distortion due to background noise. The President's room is heated and a log-fire was lit which, again, satisfied physiological needs and contributed to a relaxed atmosphere.

b) Time

After discussion with a number of fourth year medical students, we decided to offer one lunchtime (12:00-13:30) focus group and four evening (17:00-18:30) focus groups. It was felt that these timings would allow the majority of potential participants to attend, either between lectures at lunchtime or at the end of the day. In addition, 90 minutes is generally felt to be sufficient time to explore the beliefs and thoughts of participants without being too long or onerous (Asbury, 1995, Stalmeijer et al., 2014)

Email reminders were sent out to every participant one week before their scheduled focus group with information on how to gain access to the LMI. 2 days before the focus group every participant received an email to remind them of the date and time. This repeated contact is recommended to ensure that participants show up on the day (Morgan, 1995).

3) Focus Group Moderator

During his introduction, the author attempted to minimise any perceived power gradient by emphasising his own relatively recent medical studentship and the fact that he would not be involved in any manner in their assessment outwith the research project. Twohig and Putnam (2002) also warn that being a

perceived expert on the discussion topic can stifle discussion and the author therefore reinforced the idea that he was here to learn from the participants. The author also explained that everybody's thoughts were valid and that it was not the goal of the focus group to achieve some sort of consensus (Carey, 1995a).

For the first focus group, Dr Simon Watmough, who had several years' experience in qualitative methodology and focus group operation, accompanied the author (Watmough et al., 2006b, Watmough et al., 2006a). Dr Watmough was able to ensure that the focus group was moderated appropriately and gave advice for the remaining four focus groups.

The author facilitated the discussion based on the interview guide and encouraged participants to express their views without appearing judgmental (Beyea and Nicoll, 2000c). In addition the author remained aware of "social loafing" where silence by a participant may be due to agreement or disagreement with the topic under discussion or the group "consensus" (Morgan, 1988, Kitzinger, 1995).

In all focus groups there is a balance to be sought in terms of interview standardisation (i.e. how similar are the questions put to each group) and moderator involvement. Morgan (1997b) argues that when there is a strong notion of what the research question is, having relatively standardised focus groups with a higher degree of moderator involvement allows that research question to be explored more fully. The risk of this strategy is participant disengagement if they feel that they are unable to set their own agenda and/or pace. In addition, heavy-handed moderator involvement would threaten the rationale for using focus groups, instead of group interviews for example, as we were interested in the discussion between participants (Stalmeijer et al., 2014). As our research questions were relatively clear, it was felt that the questions could be standardised (with minor changes, see below) and the discussion was guided with an appropriate level of intervention from the moderator.

4) Group Cohesion

Group cohesion is influenced by the above-mentioned 3 factors: group composition, research setting and focus group moderator. By considering these 3 factors at the planning stage, we aimed to maximise participants' willingness to contribute to an open discussion. As mentioned above, we aimed for sufficient homogeneity across the groups to allow inter-group comparison but enough heterogeneity within groups to "increase the diversity and range of positions taken on [the] issues that are discussed" (Fern, 2001).

5) Group Process Factors

According to Fern (2001) there are a number of group process factors which have been posited to influence the focus group. However he claims that recent research has shown that there are only 2 which are significant: distractions and information sampling.

a) Distractions

Fern claims that when participants are involved in a discussion they are often thinking about what they would like to say next, rather than the current topic. In addition, when listening to others they may forget what they were going to say. These distractions mean that there may be fewer original ideas in focus groups. However, for our purposes focus groups were not being used to come up with new ideas but rather to discuss the thoughts of the participants regarding professionalism, teamwork and leadership. Therefore distractions, although unavoidable, should not have had a detrimental effect on our study. In addition, the solutions provided for dealing with this problem, such as interrupting whenever a new thought occurred to someone, or writing down thoughts, would have interrupted the flow of the conversation.

b) Information sampling

According to Stasser and Titus (1985) there are two types of information within a focus group: shared and unshared. Shared information is possessed by all group members, unshared information is unique to a participant (they are unclear about information that is possessed by a few, but not all, participants).

The group process will determine how much of this information will be brought to light during the discussion. Whichever type of information is more predominant is the type more likely to be discussed. Therefore during focus group planning one needs to consider which type of information one wishes to know about. For our focus group we wanted to know about the behaviours and attitudes of the participants as fourth-year medical students. We expected a large amount of shared information and a smaller degree of unshared information. The Results section confirm this prediction, with a lot of shared experiences discussed but the occasional piece of unshared information brought to light.

6) Focus Group Discussion Process

The focus groups began with the ground rules regarding confidentiality and introductions. We then used a questioning route as a questioning strategy. According to Kruger and Casey (2000) the main benefit of using a questioning route is that it forces consistency across the focus groups and therefore improves analysis.

The draft focus group questions were arrived at by reflection on the research question, as well as a literature search which encompassed both the theory of focus group questioning and articles which used focus groups as their research method (Appendix 3-1). These draft questions were scrutinised at a one-hour discussion meeting, as recommended by Krueger and Casey (2000), held with all researchers at the Centre for Excellence in Developing Professionalism (CEDP).

There was an initial question which all participants could respond to. According to Abury (1995) this:

> "not only helps emphasize the similarities between the participants, but also brings all participants into the discussion and suggests that all contributions are equally valued" (p. 417)

In addition, Carey (1995a) claims that the longer a participant is silent, the less likely they are to speak. From the initial question we followed traditional focus

group methodology by using mainly open-ended questions (Krueger and Casey, 2000) and by using a funnelling design (Morgan, 1988) of transitional questions (Morrison-Beedy et al., 2001) to move from general to more specific questions (Beyea and Nicoll, 2000c). The final questions were an attempt to determine if the notes the author had recorded reflected the discussion that had taken place (Krueger, 1997) and to determine if anything relevant had been missed (Morrison-Beedy et al., 2001).

The resulting list of questions was piloted, as recommended by Morgan (1995), on members of the same year group who were not taking part in the focus groups. These three fourth-year students had agreed to participate in this aspect of the project (out of a total of ten students who had been emailed.) Each student spent half an hour with the author discussing the focus groups questions and suggesting modifications. These discussions led to further changes in the questions. Lastly, as recommended by Côté-Arsenault and Morrison-Beedy (1999) the list of questions was reconsidered after each focus group session and changes made as necessary.

7) Focus Group Outcome

According to Fern (2001): "Whether the [focus group] outcome is a success depends on the researcher's qualitative judgment" about three outcome components: task performance effectiveness, user's reaction and group member relations.

Task performance effectiveness refers to the quantity, quality and cost of the data collection. In terms of quantity, we achieved data saturation and the quality of the final four focus groups was excellent. The cost of the data collection was reasonable both in terms of time and money.

User's reaction refers to the satisfaction of the client to the process and outcome. In our case there was no "client" but this may instead refer to the MD supervisors and panel.

Group member relations refers to how cohesive and lively the groups were and, from reading the transcripts, one can get a sense that the groups were very relaxed, lively, understanding and humorous.

Having considered our conceptual framework we move on to the methodology of data collection, transcription and analysis.

Data collection and transcription

There are a number of options for recording focus group discussions/data. These range from taking notes, either contemporaneously or retrospectively, to audio-recording to video-recording. It was felt that video-recording would be too intrusive, a sentiment echoed by Morgan (1988), while using notes exclusively would result in the loss of too much useful information. It was decided to carry out audio-recording with brief contemporaneous note-taking of key interactions, as recommended by Howatson-Jones (2007), and more extensive retrospective notes of the author's feelings and thoughts about the preceding focus group.

The first focus group was recorded on audio-tape only. Problems identified with this form of data capture meant that subsequent focus groups were recorded on two digital audio-recorders. Contemporaneous and retrospective notes were retained in order to supplement subsequent analysis. As recommended by Côté-Arsenault and Morrison-Beedy (1999) participants were made aware of the data-recording from the outset. In addition, participants were not discouraged from using first names but were informed that these would be anonymised during transcription.

The recordings were transcribed verbatim using Transcriva© (Bartas Technologies) and annotated using the symbols defined in Appendix 3-2. Although verbatim transcription is time-intensive it is also the most rigorous (Beyea and Nicoll, 2000a). However, there is a trade-off between the complexity of the transcription and its readability (Bourdieu, 1996). Therefore, we

considered the need for the transcription to be sufficiently detailed without becoming unreadable.

Framework analysis

According to Beyea and Nicoll (2000a), "The goal of analysing and interpreting data is to reduce the enormous amount of raw data that have been collected to a manageable aggregate" (p.1281). As recommended by Kitzinger and Barbour (1999), the transcriptions were read and re-read while listening to the audio recordings and a pragmatic grounded theory analysis (Melia, 1997) was carried out. Grounded theory, as a methodology, is consistent with the paradigm of critical theory and the epistemology of constructionism. Its developers, Glaser and Strauss, saw "empirical "reality"… as the ongoing interpretation of meaning produced by individuals engaged in a common project" (p.633)(Suddaby, 2006)

Although many researchers might claim to be carrying out a grounded theory approach to analysis, this assumes no *a priori* beliefs (Barbour, 2007). However, the entire organisation of focus group research including question design, sample size and sample selection depends on having some prior beliefs and these must affect the final outcome. Pragmatic grounded theory therefore acknowledges that some of the themes may be predictable from the outset, Ritchie and Spencer's (1994) "*a priori* codes", but that this must not prevent their revision or emergence of other codes (Crabtree and Miller, 1992). This approach has also been advocated by Lingard (2014) who used the term "constructivist" instead of "pragmatic" but stated that constructivist grounded theory does not "imply a process of discovery untainted by prior knowledge".

The author used NVivo8 (QSR International) to carry out the framework analysis. Analysis began at the end of the first focus group. The initial analysis was descriptive but cyclical, i.e. if a new code was identified, preceding transcripts were re-analysed to ascertain whether or not this code could be matched to additional discussion. This "first coding pass" looked for manifest content codes, described by Morgan (1997a) as "concrete things which can be immediately recognised and marked". In addition we followed a

recommendation by Kitzinger (1995) to code certain types of narrative (e.g. jokes, anecdotes) and certain types of interaction (e.g. questions, deferring to the opinion of others, changes of mind). We also looked for broad categories during this first pass. This coding was reviewed by, and discussed with, Dr Simon Watmough in order to ensure that the process had been carried out correctly.

The second coding pass is inductive rather than descriptive (Miles and Huberman, 1994), and allowed us to do three things:

1. Refine the initial codes and aggregate similar codes,
2. Expand the broad categories, using a process of constant comparison (Glaser, 1965), into more specific sub-categories and
3. Detect "divergent views" among the participants (Powell and Single, 1996). This "deviant case analysis" forced the author to rethink and refine the analysis. (Kitzinger, 1995, Seale, 1999)

Data presentation

The results of the analysis are presented below in an interpretive summary format (Morgan, 1997a). This provides a descriptive précis of the answers to the focus group guide questions followed by an interpretation in the Discussion section. This data presentation reflects a horizontal, question by question, analysis across groups (Rausch, 1997). This form of analysis is recommended for beginning moderators (Krueger, 1997) and this form of reporting is supported by Kinzey (1997). Each question is presented followed by the number of words in the transcript devoted to the discussion of this question and the number of participants involved in the discussion of that question.

In addition to the question by question report, broader themes are presented, with supportive data, using a vertical analysis within and across groups (Krueger, 1997, Rausch, 1997).

The data are presented in the form of individual quotes but also, when necessary and to show how participants interacted, as a conversation.

Results

Ethical approval

Letter for ethical approval provided as Appendix 3-3.

No illegal acts were discussed in the focus groups. Instances of unprofessional behaviour were discussed and explored within the groups. The author did not feel that a private chat with any of the participants was required, although the ethics approval made provision for a private chat with participants if this was warranted.

Conceptual framework

Group composition

Of the first group of 60 medical students asked to participate, 18 agreed to attend a focus group. Of the second group of 59, 17 agreed to attend. Appendix 3-4 details the selection process. These 35 students (18 male, 17 female) represent 10.3% of the total 4th year cohort. Table 3-2 displays the group composition on each of the focus group dates.

Table 3-2: Group composition

Date	Attendees M/F
17th Feb	3/3
20th Feb	4/1
25th Feb	2/4
26th Feb	5/1
27th Feb	4/4

The focus group literature recommends over-recruiting by 20% (Morgan, 1997b), as this is the "no-show" fraction. We recruited 35 people; 4 participants (11%) failed to attend.

Focus group moderator

The moderator explained the lack of requirement for consensus by phrases such as:

> "No that's fine I want to hear everybody's ideas, that's what I we're here to be talking about... There's no consensus I just want to get people's ideas, thoughts, that's all"

> "I'm not trying to get one single, true answer to anything. I don't want a cons… It's not a… we're not here to sort of come up with: this is the answer to that or this is the answer to this… If there's disagreement it's fine…"

Social loafing was discouraged by repeated requests for further views, at times directed at individual participants:

> "Does anybody else agree with that that is gets... it's different for every year?"
>
> "Everybody's kind of nodding would you mainly agree with that mostly agree?"
>
> "[Participant's name], anything else? That makes a professional…"
>
> "[Participant's name], any thoughts on that? What makes a good team?"

The percentage coverage of the transcript by the moderator, i.e. how much of the discussion was taken up by the moderator, after removing focus group guide questions, is shown in Table 3-3 below.

Table 3-3: Percentage coverage of transcript by moderator

Focus group	Coverage (%)
2	8.47
3	11.73
4	7.46
5	7.39

Focus group discussion process

The list of questions used with the last focus group is attached as Appendix 3-5. Beyea and Nicoll (2000b) suggest that the average number of questions for a 90-minute focus group is 12 and our interview guide had 10 questions.

Data collection and transcription

Unfortunately the sound quality from the analogue tape used for the first focus group was so poor that the discussions from the first focus group were not transcribable. Beyea and Nicoll (2000a) comment that it "may be wise to use two tape recorders" and the author took this advice for the subsequent four focus groups, using two digital tape recorders.

From re-reading the transcripts and listening to the audio recordings, it was clear that by the final focus group "data saturation" (Morgan, 1997b) had been reached, i.e. additional focus groups were unlikely to provide new data.

Data analysis and interpretation

Codes

Our final code book consisted of 274 codes and, where the title was not self-explanatory, their definitions. Examples from the code book are provided in Appendix 3-5.

Questions

We first present the results from the 9 questions.

Question 1: What comes to mind when I say the word "Professionalism"? (22 participants, 536 words)

The majority of the participants discussed professionalism in the sense of visible outward manifestations, particularly dress, rather than an internalised ethical or moral code.

> “The way that one, sort of, portrays one, it, ourself to different people or to people in public, ehm the image that they convey to others.” M1
>
> “I think about appearance like smart ((laughs)) ((laughter)) clothes and stuff.” F3
>
> “There's been more of a like more of a traditional aspect to it as well which involves kind of the way the way you dress and the way the way you speak kind of dress and behave as your grandma would… would be proud of kind of thing.” M10
>
> “You just think that you know you have to be professional you have to act in a certain manner…” M15

Although some participants did mention internal constructs such as competence and expertise.

> “When I thought about it more maybe competency comes in as well…” M4
>
> “…when I think of professionalism it's ehm having the expertise and ehm and ehm using that expertise to the sort of the best sort of fit of the situation.” M6
>
> “Yeah I think professionalism is like how the way in which you ehm project it to the patient and to the other staff ehm like that you're actually competent at what you do.” F10

Lastly, when internal constructs were mentioned, a discussion ensued in which participants argued about the importance of expertise, experience or competence and the ability to portray this. Was it more important to have expertise or more important to be able to appear confident or were both important?

> “I think f.. eh for a patient professionalism is is being confident in what you're doing [and not so much what you know to to some extent” M8
>
> “It's all well 'cos you can have like all the knowledge and all the training in the world but if you can't impart that with someone that's depending on you then it's not...” F11

"…the patient you got a good rapport with them and then he'll be: "Yeah, he's a professional." But do you actually know what you're doing?" M11

Question 2: Is the professionalism expected of medical students different from that expected of doctors and, if so, why? (15 participants, 1227 words)

The participants were unanimous that the professionalism expected of them was different from that expected of doctors. On the whole the participants felt that medical students were not expected to be as professional as doctors:

"…it's slightly accepted that you can have a life outside and you can let's say go out and get drunk and come in whereas if a doctor was to do that." F4

"…they expect us to go out and get drunk and go and party that sort of." M10

This variation in expected professionalism was felt mainly to be due to the reduced responsibility of medical students.

"Yeah cos you're not like directly responsible for the patients. You don't actually give give any care to them. You're just there to learn." M5

"You don't really have responsibility do you? Or liability." M6

"In that we don't really have many responsibilities do we?" M10

"…you're having more of a role in their care the higher up in the years you are. So I think they expect more professionalism from you." F7

Lastly, the majority of participants felt that the professionalism expected of medical students, although less than that expected of doctors, did increase through the years:

"It's... not everyone automatically has professionalism it's something that I think absolutely has increased with you know every single one of us through the years of clinical practice but it's something you develop" M1

"I mean as senior medical students we get the younger ones looking up to us. So you need to set an example from that point of view. And you need to build up the way you know you you have your your own professionalism you need to bring build that up as you go through" M2

"Which is why I don't think you've even got palliative care until fourth year because your professionalism is increasing…" F7

Question 3: If you see unprofessional behaviour by another medical student, how do you deal with that? (23 participants, 5601 words)

Most participants, who would be prepared to act on unprofessional behaviour, would speak to the transgressor in the first instance.

"I have talked to people kind of warned people where, I sound like such a killjoy, but so I have sometimes sort of anticipated maybe certain people are gonna be a bit un.. eh like unprofessional" F9

"…maybe you'd just tell them themselves that "You know this isn't appropriate."" F10

"But I I'd maybe say something to them but I wouldn't eh also I've not I've never like t-taken anything further than that." M9

None of the participants referred to a specific method for reporting unprofessional behaviour. Participants who were asked directly if they were aware of such a method replied in the negative.

"But I wouldn't go and ring faculty" F12

"So I think if you did tell someone like I don't know your tutor or someone in the medical school" F9

"If there is [a reporting system] I'm not aware of it" M13

"There probably is [a reporting system] but no-one knows about it" M15

The reasons for not using official channels for reporting behaviour ranged from fear of over-reaction by the faculty, to a sense of collegiality, to a belief that it was up to "others" to see and report unprofessional behaviour.

"I don't think I'd ever like to get a medical student chucked off a course and I'd I'd I'd feel like that's not really my place to and I'd feel like it..." M4

"Cos you really don't want to be.. like get someone in loads of trouble do you know what I mean?" F9

> "Cos you would worry about the kind of you know how far it they would you know people might the faculty might take it… And that you know just cos they've said it doesn't mean they need to get chucked off the course." M15
>
> "So if we were to report back like XXXX was saying we would actually be scared that the repercussions would be like a lot of worse than what. Vastly out of proportion to what actually happened." M14
>
> "I've heard of other doctors saying "Oh you've gotta look out for your colleagues. You've gotta remember this 'n appreciate we're in it together you know." F6
>
> "I think I'd I'd thought in past there's people who I don't think are that professional but surely the University will find that out. Maybe. I don't know." M4
>
> "I think it's di.. I think you personally you wait until a senior says [something to them." F5

Lastly, the majority of people who expressed an opinion thought that the University would not be able to identify unprofessional students in the existing set-up. Either because the students were not being monitored or because unprofessional students could "tick the boxes" when required to do so.

> "No There's nowhere to there's nowhere I think you can, you know, the way our course is run, there's no way that you could pick it up really I don't think" M2
>
> "Yeah they know what they're supposed to say… tick boxes" F3
>
> "You know we all know what we're supposed to say… tick box thing" F6
>
> "So you are pretending to be sympathetic to get a tick basically" F9

Question 4: Do you think that what we think of as "professionalism" today is different from what people would have thought of as "professionalism" 30, 40 or 50 years ago and, if so, why? (19 participants, 2426 words)

All the participants who discussed this question thought that what we think of as "professionalism" had changed. The two main changes were felt to be a focus on patient-centred care and, either as a consequence or as a concurrence, an emphasis on empathy and communication skills.

> "Doctor was right, patient didn't have a say. Whereas now it's you know flipped the other way it's patient-centred care and that's the professional way now…" M2
>
> "Whereas now it's very much patient-centred care so they decide what they want, they're given options 'n they expect that as part of part of the professionalism of the doctor…" F7
>
> "But now it's much more of a kind of relationship and it's like equal yeah, more patient choice exactly." F9
>
> "And I think that that reflects as you know how a a a doctor in kind of like the old-fashioned sort of didactic role ehm was appreciated back then compared with how a more caring ehm you know emotionally receptive doctor is in in in the modern kind of eh relationship." M7
>
> "Yeah I think there's definitely more of an emphasis on communication skills now than what there used to be" F10
>
> "Whereas now it's a bit like a bit more if you m.. make sure that you build a rappo... up a rapport with the patient and you're gettin all them communication skills in there." F3

Question 5: Think of somebody you've met or seen at work who you think is "professional" what did they do or say, how did they act, to make you think this of them? (15 participants, 1059 words)

The discussions in response to this question revolved around three main themes. The first was patient-centred care.

> "They're genuinely concerned for the patients and their views." F9
>
> "They listen and respond to patients." M2
>
> "I mean recognise their pa- the patients concerns as well." F10

The second theme was competence.

"Competence" M6

"I think havin a good routine going in, introducing themselves and then doing a proper examination thoroughly and quickly is always quite I think I think oh they they know what they're doin." F5

"I think on top of that they need to be they need to be competent so they they obviously have to have a great knowledge about what they're doing and be a good diagnostician and then be able to treat appropriately…" M4

And the third theme was external appearance/dress.

"Going back to image I guess the the they're dressed appropriately that's the bit you can jus.. you can see straight away." M3

"I suppose appearance as well. If you go in as wearin jeans or somethin like that." M5

"Or like inappropriate skirt and low tops." F5

Question 6: What do you think about "bringing the profession into disrepute"? Is that still relevant today? (25 participants, 7317 words)

Quantitatively, this question generated the most discussion both in terms of numbers of participants involved and the number of words used.

The majority of participants felt that "bringing the profession into disrepute" was still relevant today.

"I think it is because when you're working as a doctor you it's different from any other job you're working in a position of trust…" M12

"I think you definitely have to be a bit stricter with people in in that kind of position." F9

"cos you kinda take I think doctor's like a job that you kind of take home as much as anything you are still a doctor. You're not you're not just like you could be an engineer at work and then you know normal at home whatever." M15

> “But if you went out of your way to bring it into disrepute then yeah like the patient doctor trust is a sort of key part to treatment and if someone goes out of their way to break that down then yeah” M9

A minority of participants expressed disagreement with the concept of disrepute with the main thrust being a separation between private and public/home and hospital, while others queried how much reputation the profession has.

> “I think at the end of the day you've got to say well they're not letting their they you know they've got 1) they've got a right to a private life and 2) that's it's not affecting my care. And if it did affect my care that's when it becomes unprofessional and that's when it's a problem but…” M2
>
> “I don't think the fact that you do something outside of hospital and outside of work should have an effect on how you look after your patients if it's in that sort of situation.” F12
>
> “But it's like you you've become a doctor but your doctor isn't your life. Do you know what I mean? You don't have to be professional then twenty-four hours a day, seven days a week just because you you are a doctor.” F7
>
> “Say you've you've got one racist doctor and how much disres-pute has he brought the rest of doctors into and is it really worth ruining his career?” M14
>
> “I don't I don't necessarily think this but ehm or b.. sorry believe this but almost in a way that would be a situation where the GMC and sort of like doctors as a body sort of have too high an opinion of themselves” M7
>
> “And I think I I agree I think it's a GMC sort of almost thinking a bit too much of sort of..” M9

Most participants agreed that decisions around disrepute needed to be made on a case by case basis.

> “It really does depend on circumstances like you don't know why people do the things they do” F9

> "I suppose it does depend on the situation a bit doesn't it? And the specifics of it." M15
>
> "You have to see each case with its merits and then sort of see.." M5

Although much of the discussion regarding who has authority to decide what is disreputable referred to the GMC, many participants also referred to patients.

> "But in a kind of like as a patient I suppose it kind of a lot of it depends on how the patient would feel after they've been treated by someone" M15
>
> "I think these sort of things it's like it's up to the patients either to find a new doctor or forgive them and then with other sort of worse things maybe like negligence n n sort of your care of the patients. then it's up to the GMC to do something about it." M
>
> "No no I was just going to say I think I think you're right but sadly I I don't think patients see it the same way at the moment" M8
>
> "I think the problem would be gauging it on what what patients' views were and I think if you were to take a poll of patients they would say "I don't want that to happen" and that's why it is unprofessional." M4

Question 7: What makes a good/bad teamworker? (21 participants, 1634 words)

A characteristic of a good teamworker discussed in every focus group was role clarity.

> "and knowing what knowing what each person does in the job. Is always helpful and knowing what you're supposed to do and what somebody else is supposed to do." F5
>
> "Knowing your role within the team that you're part of." M5
>
> "I think also it helps if everyone has a defined role." F1
>
> "Cos that's something I mean to me that is it's always amazing to see you know how everyone just jumps in they've got their role…" F11

The participants also mentioned that a good teamworker communicates well.

> "Good communication. in any team. Even in sport like eh if you don't have communication it all breaks down" M12

"People being open and verbal is a good team rather than eh bottling things up or not voicing opinions that you find important" M6

"I think communication between the members is somethin that's probably quite simple but very very important." M1

The need for respect of other teamworkers was another frequently mentioned behaviour.

"Yeah respect for others as well." M6

"Usually I think there's there's a healthy respect for each other ehm within it and that goes with with the respect for the the role that that they play" M4

"Ehm and it's about respecting others ehm…" M1

"Like particularly like you've seen in like MDT meetings everyone has a you know everyone's role's respected…" F1

Lastly, the participants discussed the need for teamworkers to be willing to contribute to the team:

"And then the opposite as well not taking back standing back and sort of I mean like always behind and ehm I'll just stand back here and you can do it. That just doesn't work." F5

"Not working with other people but working on your own almost…" F3

"…you're all kind of in it together and you help each other out. I think that would be good." F6

"I think each everyone should contribute equally if like just one percent contributes and then the rest like doesn't do anything it wouldn't work at all." F2

Question 8: What makes a good/bad leader? (23 participants, 2504 words)

Many of the participants stated that the leader should be inclusive rather than dictatorial.

"I don't think they should be necessarily being a dictator and telling 'em exactly what they should be doing "This is it!"" F5

> "need to be in the middle whereby they have that authority and people respect their decisions but they're not they're not too authoritative ehm so that people just dislike them and don't enjoy their job" M4

At the same time, many participants felt the leader needed to be able and willing to challenge poor performance.

> "Also have the ability of reining them back in as well like." M5
>
> "And if someone is not doin' their job to have the confidence to say: "You're not pullin your weight." 'Cos you know it's all very well being a nice team leader and everyone's like "Oh he's such a nice guy" but if you can't tell someone "You're not doin' your job" then what sort of leader are you really?" M2
>
> "Because I mean like I work you know in a shop and our boss is the leader. If you're not doin your sales, she will tell you you're not doin your sales." F1

Participants discussed the balance between being the leader and yet remaining part of the team.

> "…they shouldn't think they're better than everyone else that's just that is their job as part of the team. It's just it's no different to any other job on in that team." M3
>
> "But at the end of the day that doesn't make them a a better person or a bigger person. They're on the same level as everyone else… They're they're only above you in sort of a hierarchical sense not a... person sort of sense. M2
>
> "A leader has to be someone that everyone agrees should be the leader and not just TAKES the role on because they think they should" F9

Other attributes referred to included: decisiveness, situation awareness, experience and confidence.

Question 9: "Is there anything else you want to talk about with regard to professionalism?"

Due to the catch-all nature of the question, the responses were diverse and could not be classified into themes. Some focus groups returned to talking about disrepute, this discussion was included in the analysis for Question 6 above. Others talked about bullying, which is covered in the "hidden curriculum" theme below and one focus group talked about assessing professionalism, which is discussed in Theme 2 below.

Themes

We identified six themes after analysing and comparing the focus group discussions. Each theme arose during the inductive, second pass of coding. The discussions were listened to and re-read to look for both supporting and refuting evidence.

Theme 1: "Acting" versus "being" professional

Given the topic of the focus groups, participants used the word "professionalism" and its derivatives "professional", "professionally" on a number of occasions. One of the themes which emerged from the analysis across all focus groups was the dichotomy between "being" and "acting" professional. Examples include:

1) Acting

"It's kind of of how you conduct yourself around patients and around your colleagues too and just how you act" M12

"Yeah I think professionalism is like how the way in which you ehm project it to the patient and to the other staff..." F10

"I think I think professionalism is making is trying to make yourself try to inspire respect from a patient trying to make yourself portray yourself in a way that ehm" M8

"I think it kinda comes down to sort of the most appropriate behaviour in any given situation isn't it?" M7

"The way that one, sort of, portrays one, it, ourself to different people or to people in public, ehm the image that they convey to others." M1

2) Being

"They introduced this thing saying ehmm you know medical students should be professional outside of hospital and clinical care etcetera" M2

"I mean I don't think anyone can be professional a hundred percent of the time" M1

"I think that that's the situation where because he is a doctor and he is professional, he can't do it." F3

"I think you're expected to become more and more professional." F7

"Also being professional amongst colleagues as well would make teamwork 'n the team work better obviously with the whole team with physios, OTs, nurses and doctors and everything." M13

Theme 2: The hidden curriculum

While the official curriculum tells the undergraduates what should happen, the hidden curriculum shows them what actually happens.

"But I find that like because like there's some doctors are higher like say consultant they can act less professional some like I've seen doctors less less professional professionally but they get away with it cos their the ranking." F8

"Some doctors you can just kind of tell just go "Oh that must be terrible for you." Walk back and just like "Watching the football tonight?" (Laughter)" M15

"So you'd think somebody who's being a who would be an example to us, a consultant, who's actually teaching us. And then are you actually going by his behaviour, doing what he always does? Or are you gonna..." F6

The hidden curriculum also encompasses belittlement, humiliation and disrespect.

"there are definitely older consultants who ehm think that we're kind of lesser because of our.. because of the way that we're taught and.. well not even the way we're taught but the way we kind of like approach the training and eh you know we're kind of told not infrequently that we you know that it's it's rubbish, it's not effective that we're not basically not as

good as they are and I suppose that comes into professionalism as well" M8

"I had a a mate who got named after a colostomy bag ((laughter)) by a consultant." M2

"And ehm I had em a dermatologist and he went: "So you do... Wanna do Medicine do ya?" I went "Uh-huh" He went: "You know women are ruining the NHS." F1

"Someone like someone who's in our hospital group had a German name and two different patients he called her Nazi cos she had a German name." F9

"Yeah and started asking he started asking questions and then eh he started asking questions I answered one and he turned to me and just said "Oh you look like a bin man.""M

"The most recent thing that happened to me was I ehm I got an SSM regraded. So I sent an email saying to someone ehm "When's this going to be reflected on my on my Sp.. Spider transcript and didn't get a reply. I waited a few weeks. I sent in another email saying "Did you get my email?" and then they just sent back "Oh check Spider, it's done mate." And I was like oh thanks for being courteous and replying back to me in the first place and just letting me know. And that's just one thing that happened recently. The-the-there's a string of things that happened over the time that we've been at medical school. Lots of times when you haven't been treated professionally. It's as if we're like ehm maybe maybe yeah it's too strong a word but we aren't treated the same as how they treat other people." M

The participants provided a number of rationales for why unprofessional behaviour by their superiors was acceptable:

"Ehm and I think the other thing is looking at the consequences for that person so if I were to report someone for something serious who I thought was still doing a job as a as a consultant and still probably saving a hell of a lot people despite doing something unprofessional I think to think maybe they might get sacked as a result of that" M4

"Ehm and it's almost that respect even for the most horrible of consultants there's still that that respect and if it ca.. it's somethin it it seems to go throughout sort of eh medical school and then into ehm into doin it as a job." M4

"I I sometimes though almost expect to be [grilled and expect to be bullied and it... 'Cos we're at the bottom of the bottom we're below the patients in hospital. ((laughter))" M2

Or if not acceptable then not confronted:

"Most medical students wouldn't say anything but because you're so scared of what's gonna happen in your future career. like who's going to be employing you because like they're basic.. they could be your boss like that's gonna be doing your interviewing. You don't wanna like kind of rock any boats while you're there." M

"It does sound like a class example of someone who's really high up and therefore untouchable." M

"I need my book signed at the end of that session ((Laughter)) and and I wasn't willing to sacrifice that and and I just thought I'm not going to achieve anything by this he's not going to change his ways" M4

"Then again you do have to have like a bit of confidence to go up to a consultant and go "Actually you know what, what you've just done I don't agree with." And I wouldn't do it I don't think." M

"It's about respect and you've gotta respect your elders I mean in Medicine there's a big culture of traditional manners, respect, respectin your elders… I mean you know sometimes it's they know the real thing but experience might have taught them to cut corners [so when we see it from our naive unexperienced eyes it looks like a lot worse than it potentially could be from their perspective" F6

Theme 3: The rumour mill

There were discussions across all focus groups where reference was made to stories regarding unprofessional behaviour:

“You hear stories about boys about students getting pulled up for [things on Facebook and things and whether they're true or not I don't know but you hear all these horror stories about someone's pic shows this and faculty get hold of it and stuff like stuff like this.” M3

“Yeah I think the whole whistleblowing ehmm you know 'cos I remember ages ago there was someone can't remember which hospital but basically they were taking like a consultant to the GMC about bullying and one of the tutors said "He's basically creat.. committing career suicide by doing that." And it's just so messages like that” F1

“But I think I mean goin through the years again these things get distorted over time. But there are a number of stories of what some medical students have managed to get away with and and you look and you think.. And then you also I don't know you look at faculty from in a maybe a slightly cynical way and you think that you know : "They could've acted on that. They didn't." And then you think "Why?" and I think and then I don't know you start thinking "Oh well they ploughed so much more money into you that they think...” M1

“Cos you might think like you know it could just be a flippant comment that when you actually say to them about it they go "Yeah I actually feel really bad about that." And that you know just cos they've said it doesn't mean they need to get chucked off the course.”M15
“Yeah (laughs)” M14
“Yeah (laughs)” F9
“And then and you can almost see the faculty doing that.”M15
“Yeah” M13
“Cos you hear about stories of... and you kind of think well” M13

“I mean some stories you hear I mean about hypothetical situations like that one (laughter) and you just think "How on earth do people get away with that?" M

There were also somewhat more factual references, such as:

> “I know someone who ehm did an SSM and they wanted to dispute the mark but the person that was the convenor was also sitting on the moderating board and so the person said to him "I'm I'm gonna to sort of appeal it" and he goes "Don't bother." (Laughter)

> “I have a friend who waited 18 months to have her SSM regraded cos eh some person, no names, kept on eh losing the paperwork. That's ridiculous you know it's paperwork it's not hard it's you job. You know it's just so annoying.” M

> “there was that that student who you know the guy that the breast surgeon who went like that "Whey" on some girl's breasts when they were asleep under anaesthetic. The med, it was the medical student not any not any of the team not the nurse not the SpR anything like that it was the medical student that was just like "Hang on a minute" (laughs) "You can't just do that.”” M

> “Yeah cos there was a situation last year where quite a few people kind of people made example of made an example of over a Facebook page. And it was kind of like... everyone was kind of like oh after that happened there was like real paranoia (laughs)” F9

Theme 4: In the eye of the beholder

In focus group analysis one is advised to examine areas of tension (Barbour, 2007). Every focus group discussed professionalism in terms of being an objective/subjective concept. Some participants felt professionalism was subjective, while others disagreed. At times it is the interactions in focus groups that are most revealing (Asbury, 1995, Rapley, 2007) and therefore conversations from two focus groups are provided below:

> “I think it's professionalism is behaving in a way that the patient you know would like you to behave. So by definition it's it is subjective

because it's how patients perceive you so there's always going to be some patients who want you to act like in a very sort of strict way." M8

"But is it is it though? That's the thing is it is it really ehm the way patients percei.. I think wha what the problem here is that we don't actually have a working definition of what professionalism is. And personally I think there's there's a couple of eh a couple of areas to it some of which may have clear boundaries although which others may not… So I think professionalism kind of encompasses all of these things not just it's not just the way your patients are perceiving you but also the way you act in medical practice as well." M11

"It's it's it's adaptive though on one part isn't it? I mean there's certain elements that are obviously kind of eh constrained by by law obviously. There there there are certain behaviours that wouldn't wouldn't be allowed because they you know they'd be like I dunno like sexual harassment or something like that. Things that you wouldn't couldn't possibly do by law. But there's also like things that are adaptive that like you know you that you would behave a diff.. a certain way ehm f.. to like an elderly member of the public that you obviously wouldn't with with a younger person because ehm in order to to to build a rapport with that person… To it's kind of it's it's the most appropriate thing in that situation so in that way it is subjective." M7

"I don't know it's kinda like ehmmm. it's not really a a definition where you could say "That's unprofessional" it's a an opinion, so some what one one behaviour to upon one person might look unprofessional but another person might say:" Well, no, I don't think that was unprofessional." M2

"Yeah like I mean there's been a few times when you have been on the wards when you do feel a little bit uncomfortable. Like I remember you know in a hospital when I'd seen a Caesarean section and they were basically discussing about what the baby looked like and basically what syndrome it must have to be that ugly (nervous laugh) And I was just thinking sometimes you know the mums can still hear you know when

they're under anaesthetic and I'm just standing there cringeing. But I wouldn't you know do that but other people you know think that's acceptable. But that's the way it is. Don't know. Yeah but...” F1

“But certain characteristics I mean they're always unprofessional like rudeness, arrogance, things like that, they're always deemed unprofessional.” M3

“It's about gaugeing the situation as well so, it is it is subjective as XXX said to begin with ehmm... and so with certain patients you can act in certain ways, you can be more brash and you can be more straight to the point and perhaps less caring if that's if that's what you think they they want. So sometimes th.. what they want to know is just they'll just want to know the facts and they'll want to be told what your opinion is. Ehhmmmm So so it's it's not a fixed thing either it depends on the situation.” M4

Theme 5: The language of professionalism

In all discussions, across all focus groups, words which encapsulate a set of ideas in medical ethics, such as non-maleficence, beneficence and justice are never used. The word autonomy is used once:

“So I think patients' patients' views have changed and and they now have that autonomy that is talked about so much in in terms of ethics and so I think that's that's had a big effect on on why professionalism has changed.” M4

Specific GMC documentation, which sets the professional standards which medical students are expected to abide by, is referred to once:

“Well there's there are guidelines aren't there. "Duties of a doctor" 'n GMC guidelines.” M6

There are a number of references to the GMC and its guidelines.

“I think perhaps that's maybe going back to what we said before where GMC guidelines come in… I think maybe the GMC guidelines help bring people into a line.” M4

> "I think probably it comes back from from things like complaints and and then realising that if [the GMC] don't set guidelines then then doctors can defend their their actions." M4
>
> "'cos I'm like one one of the thing that I I remember reading that the GMC was starting to say... I don't know correct me if I'm wrong but if they.. They introduced this thing saying ehmm you know medical students should be professional outside of hospital and clinical care etcetera and you know if you can't do that then you can't qualify it's what's expected of a doctor." M2
>
> "Yeah it's 'cos throughout the course we're taught about kind of ideals about you know professionalism and stuff and there's all guidelines…" F1

Lastly, Appendix 3-10 maps some relevant quotes to the principles of professionalism referred to in "Tomorrow's Doctors" (General Medical Council, 2003).

Theme 6: The gender of language

The majority of the participants use the first or second person singular or the third person plural when discussing "a professional":

> "The way that one, sort of, portrays one, it, ourself to different people…" M1
>
> "…someone who takes into consideration the dignity of the patient but also you know lets themselves be a bit human… And I think the best professionals are someone who doesn't just do it by the book" F1
>
> "I I don't think so. I think I think professionalism is making is trying to make yourself try to inspire respect from a patient trying to make yourself portray yourself in a way that ehm" M8
>
> "I think on top of that they need to be they need to be competent" M4
>
> "I think havin a good routine going in, introducing themselves…" F5
>
> ""cos like the doctor's got more patients in their like lives in their hands than the medical student at that present time. So if they do something wrong" F3

"To be honest I think I I mean when I think of professionalism what comes to my mind is just a guy or a girl who knows what he or she is doing." M11

"I think a lot of it is to do with communication skills and just how how just how you put yourself across and things like that." M13

At times the second person masculine pronoun is used or the doctor is identified as male in other ways:

"So because they're like top consultant or whatever the patient will say "That's good" or he's a consultant so he can get away with it." F8

"Right, he's in secondary school and he's thinking he wants to be a doctor but he's not acting how a doctor would behave" M2

"…if you're on holiday do you have to walk around with a shirt and tie just because you are a doctor." M3

""You know we can't be seen doing these things you know we're gentlemen of the profession or whatever" M2

"'Cos you know it's all very well being a nice team leader and everyone's like "Oh he's such a nice guy" M2

"Even if even if ehm he is the best surgeon in the world and he's very careful and meticulous" M6

"One little misdemeanour could mean this doctor could potentially lose his career and his life." F6

"Like I picture like an old guy in a suit with glasses like sat behind a desk" F7

"Because like say the example with the doctor smoking weed you say that but then like 10% of his patients might smoke weed anyway." F7

"For example a surgeon and I keep going to the surgeon (quiet laughter) but let's say a surgeon who is a who knows what he's doing okay? He's like you know really really good at what he does right?" M11

"Say if a doctor was an alcoholic although never drank at work and never turned up to work drunk. And the GMC found out, the hospital found out then he would need to prove that he was never.." M9

Only once is the second person feminine pronoun used when referring to a "non-specific" professional:

> "Oh she left, so and so left because so and so reported her." F6

An example of how the moderator avoided the use of third person pronouns but the discussion still moved on to using "he", is provided below:

> "And if you just think about leadership then ehm cos people say leaders have to be good communicators and there's.. What but what makes a what makes the a leader different from a teamworker. What's the you know [what makes a good leader?" Mod

> "Someone who can someone who can kinda inspire and motivate eh" M10

> "And he's quite diplomatic, kinda able to see things from many like view viewpoints." F8

Discussion

Conceptual framework

Group composition

Our 11% no-show rate is small when compared to the literature. It is unclear what the cause of this may be. The frequent reminders to attend or the incentive of potentially winning an mp3 player may have played a part. However, it may also reflect the professional behaviour of these 4th-year undergraduates.

As detailed in the Methods, we did not divide the focus groups according to gender, socio-economic background or ethnicity because we did not want to create artificial groups and we felt that the conversation would be more natural if the groups were mixed. Although the success of these compositions is open to debate, certainly different arrangements would have led to different discussions, we feel that the groups, as organised, provided a fertile milieu, In addition, the effect of gender and gender-focused language is analysed in Theme 6: The gender of language.

Focus group moderator

The author was the moderator for all five focus groups, with Dr Watmough supporting and providing feedback on the facilitation of the first group. Although the author had spent considerable time facilitating debriefs in the simulator environment, he was not an expert. Albrecht et al. (1993) emphasize the importance of the moderator's experience, communication competence and communication style. Sim (1998) states: "The skills and attributes of the moderator… will exert a powerful influence on the quality of the data collected in a focus group" (p.347). It is therefore possible that the author's lack of experience in moderating focus groups will have either affected the discussion or resulted in lost opportunities to develop the discussion. However, Morgan (1995) argues: "my experience has been that focus groups are relatively robust with regard to moderator problems" (p.521).

In terms of moderator involvement, Sim (1998) refers to Hague (1993) when he states that: "In terms of overall input from the moderator, …this should constitute between 5% and 10% of the resulting transcript" (p.347). The moderator involvement in these focus group discussions between questions ranged from 7.39-11.73%. Therefore it seems that, in terms of moderator involvement at least, the moderator did not dominate the discussions.

Group cohesion

Both Asbury (1995) and Carey (1994) counsel us not to overlook the effect of the group. Although one could provide quantitative data, such as percentage of time each participant spoke, or the number of laughs or interruptions, these would not validate the process. We felt that there was sufficient group cohesion for the participants to have a fruitful conversation but not so much that there was no disagreement.

Kitzinger (1994, 1995) emphasises the importance of group interactions and we have provided a section of dialogue when this was thought to be relevant. However, we are also aware that the single quotations provided did not occur in a vacuum, and that these were not interviews. We would argue that even in cases when a single quotation is used, we acknowledge that this occurred during a discussion.

Focus group discussion process

Although the confidentiality of the focus groups was impressed upon the participants both in writing and verbally at the beginning of every group, we concur with Bloor et al. (2001) who state:

> "assurances of confidentiality on the part of the researcher are limited… information is shared among members of the group over whom the researcher has little control" (p.25-26)

Unfortunately this is, in some ways, an inherent flaw of this methodology, which cannot be overcome. The analysis does not suggest that participants felt restricted by this knowledge, and the topics of conversation were not of an intimate nature. However, we cannot guarantee that participants did not withhold salient information for fear of it being more widely known.

Data collection and transcription

Data collection

A limitation of our study is that the author was the sole moderator. Sim (1998) states that: "Written notes are better taken by a co-researcher than by the moderator him- or herself". Unfortunately this was not feasible and it is therefore possible that a co-researcher would have made more detailed or insightful notes, and/or that participants were distracted by the notes that the author made.

Data transcription

Poland and Pederson (1998) inform us that "transcription is a transformative process, taking live conversation and changing it into a textual representation of talk" (p.302), such that "even so-called verbatim transcripts, are necessarily only partial accounts of the original interactions" (p.302). There is therefore no single, "true" transcription . Increasing the amount of detail to include lengths of pauses, in-taking of breath, inflections, etc. increases the richness of the text but makes it more difficult to read (Bourdieu, 1996). It was felt that the level of detail in our transcription was sufficient to explore the beliefs and attitudes surrounding professionalism, teamwork and leadership. However, as with data analysis and interpretation (see below) it is indisputable that the transcription is unique to the researcher and therefore is open to criticism.

Data analysis and interpretation

Powell and Single (1996) state: "The process of analyzing results is the least agreed upon and the least developed part of focus group methodology" (p.502). In addition, as with the transcription, the analysis is particular to the author. As Patton (2002) states:

> "Qualitative analysis transforms data into findings. No formula exists for that transformation. Guidance, yes. But no recipe. Direction can and will

be offered, but the final destination remains unique for each enquirer, known only when – and if – arrived at." (p.432)

The analysis and interpretation are therefore only one of many possible, and conclusions drawn from the research need to be seen in this light. However, as Rapley (2007) implies, this does not mean that the analysis can come "out of thin air" or be based on "a vague hunch". It is the researcher's responsibility to show how the data were analyzed and interpreted and that the conclusions were logically drawn from them.

Quantitative data

In her paper, Asbury (1995) cautions: "Do not treat qualitative data as if it were quantitative" (p.418) and Kitzinger (1995) states: "In general, it is not appropriate to give percentages in reports of focus group data" (p.301). However, in their paper "Rigour and qualitative research" Mays and Pope (1995) suggest asking the question: "Did the investigator make use of quantitative evidence to test qualitative conclusions where appropriate?" (p.111)

There are two ideological camps in qualitative research. The first believes that qualitative data can be supported by and support other data, in a process of "triangulation" for example (Seale, 1999). The other camp believes that the very nature and subjectivity of qualitative data makes quantitative comparisons meaningless (Bloor, 1997).

In the Results section, we provided the number of words in the transcript devoted to the discussion of a particular question and the number of participants involved. This is not to suggest that this is the sole indicator of the importance which the participants attributed to that question. We appreciate that the tone of voice, spontaneous expression of views, the amount of disagreement, etc. are also important and these are referred to within the body of the results. The number of words and participants involved may however *suggest* how important a given topic was to the participants.

Codes

According to Gorden (1992, quoted in Carey et al. (1996)) "a useful set of codes should be all-inclusive and mutually exclusive" (p.2). There is no hard and fast rule regarding the number of codes. Carey et al. obtained a total of 171 codes, while we obtained 274. In fact, it is the quality of the codes which is more important, both in terms of how they were arrived at and how they were used to develop the themes. If the latter are considered to impart a new or deeper understanding of the topics addressed then the codes will have proved successful.

Consensus

As stated in the Methods, it is definitely not an objective of a focus group to obtain a consensus. Indeed it is often the disagreements which expose the beliefs and attitudes underpinning group norms and therefore provide us with the richest data. (Kitzinger, 1995, Smithson, 2000). These exceptions also force the researcher to reconsider emerging themes and theories, so-called "deviant case analysis" (Kitzinger, 1995, Barbour, 2005).

Some promote the use of focus groups as a way of providing "safety in numbers", allowing recalcitrant individuals to speak out (Lederman, 1983). This view is held by Watts and Ebbutt (1987), who found that people were more willing to divulge criticism in focus groups than in one-to-one interviews. Others warn that:

- focus groups can suggest consensus where there is none (Asbury, 1995, Barbour, 2005)
- apparent consensus is an emergent property of the focus group method (Sim, 1998)
- focus groups can lead to "groupthink", a consensus-seeking tendency in an effort to preserve group harmony (Griffin, 1997).
- focus groups can be dominated by one or more individuals and the "consensus" is their opinions (Smithson, 2000)

In addition, Sim (1998), quoting Turner (1991), claims that when there is consensus within a group it may be exaggerated through a "group polarization

effect". The consensus converges on the positive or negative end of the spectrum and is amplified.

However, while it may be misguided to look for consensus within a group, it may be possible to identify consensus between groups, particularly if an issue has arisen and been dealt with (in terms of content and discussion) similarly across a number of groups (Sim, 1998). We used this approach both with our analysis of the Questions responses and our Themes.

Questions

Question 1: What comes to mind when I say the word "Professionalism"?

Wear and Kuczewski (2004) explain that the linguistic sign is made up of a sound image and a concept. For example the sound image "bucket" brings to mind the "bucket" concept. It is much more difficult to carry out the same procedure with abstract concepts, such as "professionalism". The sound image does not in this case evoke a concrete concept but rather a number of associated concepts. It is these concepts which the medical students explored.

In his paper "What medical students know about professionalism", Hafferty (2002) concludes "not a great deal" (p.396). Jha et al. (2006) interview study of a range of healthcare recipients and healthcare personnel, including medical students, found that professionalism was expressed either as a conceptual or as a behavioural component, although the relative incidence of each was not declared. The fact that the majority of participants in our study did not talk about professionalism as a conceptual/internal construct, but rather as a way of "acting", and in particular the emphasis placed on dress may suggest that "professionalism" as a field of study has either not been embedded, or is not considered sufficiently important, within the curriculum. An argument against this interpretation is the way the question was phrased, as an open question, rather than: "Please list the elements of professionalism". Alternatively, it may simply be easier to think and talk about behaviour and dress than about beliefs and attitudes. Our findings are supported by a survey by Morihara et al. (2013)

of medical students at the University of Hawaii. In response to the question "How would you define 'professionalism'?", they found that students defined it in terms of behaviour (46%), showing respect (36%) or possessing integrity/honour (33%).

Interestingly, none of the participants referred to professionalism as being imposed on them, which was one of the main conclusions of an Australian medical undergraduate focus group study by Cuesta-Briand et al. (2014). Their study included medical students in their 4th, 5th and 6th years of study. It is unclear why our participants did not express this view. It is possible that the Australian curriculum, which includes professionalism lectures, a personal and professional development mentor and formal assessment through a reflective portfolio and an ethics essay, makes more overt professionalism-related demands of its students.

The participants seemed to be aware of the disconnect between "acting professional" and "being professional". They appreciated that someone can act professional and, as this is one of the main ways that professionalism is assessed, therefore be considered to be professional without having the competence, expertise or experience which one would expect from a professional. This is explored further below in Theme 1.

Question 2: Is the professionalism expected of medical students different from that expected of doctors and, if so, why?

In their views that the professionalism expected of medical students differs from that of doctors and that professionalism increases through the years of medical school, the participants are supporting the proto-professionalism concept described by Hilton and Slotnick (2005). These views are also in line with the GMC's guidance, which refers to the development of suitable attitudes and behaviour (General Medical Council, 2003).

Where these focus groups may expand current knowledge is with the idea that the majority of participants see the increase in professionalism occurring alongside, or as a result of, an increase in responsibility. One participant, M6, said:

> "I I think I've found when when the department or wherever you're working makes you part of the team and gives you things to do and says eh: "Can you clerk in this patient?" or whatever you sort of feel more like professional."

Although responsibility is referred to by Hilton and Slotnick (2005) they do not make the link that increased responsibility may cultivate an increase in professionalism.

Question 3: If you see unprofessional behaviour by another medical student, how do you deal with that?

The fact that most participants, who were willing to challenge the unprofessional behaviour of another medical student, would speak to the offender in the first instance complies with advice from the GMC in *Good Medical Practice* (2006a): "You should challenge colleagues if their behaviour does not comply with this guidance" (p.10). This finding agrees with a survey of senior medical educators in the UK who decided that "Challenge the person about the behaviour/attitude" was the correct response in 68% of cases of unprofessional behaviour/attitude (Roff and Dherwani, 2011). (Other possible responses were: "Discuss with peers to find way of addressing" (12%), "Report the behaviour/attitude to more senior person without trying to take action" (12%) and "Ignore" (9%)).

It is highly likely that the participants have been told how to report unprofessional behaviour, perhaps even repeatedly told, and that details can be found in the various student handbooks and on the student website. However, the lack of knowledge displayed by the participants suggests that this information is not being relayed in a memorable manner, or not being reinforced sufficiently.

Participants were, on the whole, unwilling to report unprofessional behaviour using official channels. This may be, in part, because they don't know what these channels are. However, fear of over-reaction by faculty resulting in student dismissal, whether this is a genuine or imagined consequence, seems to be a major barrier. Additional focus group discussion around the role of "the faculty" is referred to in the Theme 3 "The rumour mill".

Question 4: Do you think that what we think of as "professionalism" today is different from what people would have thought of as "professionalism" 30, 40 or 50 years ago and, if so, why?

The participants were unanimous in their belief that "professionalism" had changed, a concept that is supported by an analysis of key articles on the assessment of professionalism (Hodges et al., 2011). This unanimity must be tempered by the knowledge that none of the participants were around 30, 40, or 50 years ago and therefore the discussion was based on "popular" beliefs, notions, or observations of senior doctors. The participants felt that the major change was one of increased patient-centeredness.

In the first edition of Tomorrow's Doctors (1993), the GMC stated: "The relationship between doctor and patient has changed and there is a clear duty on the doctor to be able and willing to communicate effectively…" (p.4). It may be reassuring that the participants seemed to understand and appreciate this change, which is in contrast to a survey by Gillespie et al. (2004) who found that the attitude of healthcare staff was a barrier to patient-centred care. In addition, the appreciation that not all patients want patient-centred healthcare and would prefer to either not know everything about their condition or would prefer a health professional to make a decision based on their best judgment, is supported by the literature (Little et al., 2001).

Question 5: Think of somebody you've met or seen at work who you think is "professional" what did they do or say, how did they act, to make you think this of them?

The discussions surrounding this question were, perhaps understandably, similar in nature to the discussions surrounding Question 1: "What comes to mind when I say the word 'Professionalism'?" The intent of Question 5 was to approach the same subject from a different angle, asking the participants to think about a specific person rather than in the abstract sense of Question 1. We wondered whether this would reveal any differences between abstract and concrete visualisation.

As in Question 1, the participants mentioned style of dress and competence. However, Question 1 did not see the emergence of patient-centred care as a theme. It may be that the addition of patient-centred care was in response to the discussion around Question 4, such that this aspect of professionalism is now in the participants' minds. Alternatively it may be that when the participants considered professionalism in the abstract they did not construct a doctor-patient mental model, but rather a doctor-undergraduate model or a "doctor in isolation" model. Only when they are asked to think of specific professional individuals are they then able to bring to mind the doctor-patient model and see how his adds to the professionalism construct.

Question 6: What do you think about "bringing the profession into disrepute"? Is that still relevant today?

The majority of participants felt that the concept of disrepute was still relevant, which concurred with judicial and legislative opinion. Even though the legislation governing the GMC does not refer to a duty to prevent disrepute, a report from the Law Commission (2012) reaffirms this role:

> "…the courts and in practice the regulators have long recognised that the need to maintain confidence has an important role to play in regulating health and social care professionals" (p.43)

The minority of participants who queried the concept of disrepute did so in terms of either whether it still applied outside of direct professional practice or the extent to which it applied. This finding correlates with a questionnaire study of Canadian undergraduates by Ross et al. (2013) which revealed three themes:

"free time is private time", "professionalism is unrealistic as a way of life" and "professionalism should be a way of life". The GMC and the courts in the UK are clear that a professional's private life may be subject to scrutiny and censure if it affects their practice and/or affects the standing of the profession. In a Court of Appeal case (Bolton v The Law Society (Bolton v The Law Society, 1993), referred to in a case involving the GMC (Gupta v General Medical Council (Gupta v General Medical Council, 2002)), the Master of the Rolls stated:

> "The reputation of the profession is more important than the fortunes of any individual member. Membership of a profession brings many benefits, but that is part of the price."

The queries around disrepute may also be a sign of the proto-professionalism of medical undergraduates referred to in Question 2, i.e. the professional persona which accepts both the benefits and limitations imposed by the legal and regulatory framework has not yet full developed. The GMC does consider mitigating circumstances on a case-by-case basis, which the majority of participants agreed with.

Those who queried the amount of reputation that the profession possessed may have been unaware of the polls which consistently place doctors among the most trusted of professionals (Ipsos MORI, 2011). Although Cohen (2006) states: "Evidence exists that public trust is waning…" this is contested by the opinion polls. In 2009, the founder of Ipsos MORI was quoted as saying:

> "It is a media myth that people are losing trust generally, and specifically that they are losing trust in doctors. In 1983, 82 per cent said they trusted doctors to tell the truth; now this is up ten points, to 92 per cent." (Smith, 2009)

It is unclear why these participants did not feel that the profession has a reputation to defend. It may be that they have seen a degree of unprofessionalism which has coloured their perception of the reputation of the medical professionalism (see "The hidden curriculum" theme below)

Lastly, the discussion around who decides what is disreputable behaviour was thought-provoking. The Registrar of the GMC makes an initial decision about whether a case should proceed for investigation or adjudication. The adjudication is carried out by a Fitness to Practice panel of the Medical Practitioners Tribunal Service (MPTS). This panel consists of medical and non-medical members as well as a legal assessor who advises on points of law. Panel decisions can be appealed to the High Court. It may therefore be a legitimate concern expressed by the medical students that "the public" is not involved in decisions regarding disreputable behaviour. In some instances, doctors have had support from their patients in professional misconduct cases (Evening Gazette, 2003, The Journal, 2006, Echo, 2010) and it may be incumbent upon the GMC and the MPTS to take into account such public support. In part because the understanding of what is and isn't professional depends, to some extent, on the culture in which the doctor is practising.

Question 7: What makes a good/bad teamworker?

There was less discussion across all focus groups in response to this question than some of the preceding questions. It is possible that this may be because there was a greater consensus regarding the characteristics of a good teamworker or because of the way the question was framed.

Role clarity was mentioned by all focus groups. This is undoubtedly an important prerequisite of a good teamworker. A possible reason for the prevalence of this response is that medical undergraduates frequently do not have role clarity. As O'Sullivan and McKimm (2011b) state: "It is not always easy for medical students and junior doctors to see where they might fit into the large bureaucracy of the NHS…" (p.347). This idea of not knowing their place is supported by statements made by a couple of the participants:

> "…rather than just being someone who's just standin around on the side, gettin in the way making a crowd on the ward round" F3
>
> "You know kind of w-w-w-when you first your first day on the ward in like second year you don't have a clue who's doing what, what goes

where. How to how to you know address (laughs) somebody on the ward." M9

When referring to the need for communication as a behaviour for a good teamworker, the participants are siding with accepted knowledge regarding well-performing teams. Poor communication is cited as the most common reason for medical error (Sutcliffe et al., 2004, O'Daniel and Rosenstein, 2008). Unfortunately we did not delve deeper into which particular aspects of communication the participants felt were important.

When the participants refer to the need for respect for one another, it would seem that this is an important requirement of a good teamworker. The prominence afforded to this characteristic may be because the participants feel that they, as medical undergraduates, are not sufficiently respected (see Theme 2, "The hidden curriculum", below)

Lastly, the need for teamworkers to contribute to the team is another accepted requirement of a well-performing team. Sharing the workload means that individual team members are not overwhelmed and the ability to distribute workload is an accepted benefit of teamworking (Ellis et al., 2003).

Question 8: What makes a good/bad leader?

The majority of the discussion regarding good and bad leadership centred around the inter-personal skills of the leader. The participants thought a good leader needed to be part of the team and to have a leadership style which was more democratic than dictatorial. These ideas correlate with the relational, as opposed to transactional, leadership style discussed by Cummings et al. (2010) and the post-heroic leadership style referred to by Alimo-Metcalfe and Alban-Metcalfe (2006). This may suggest that the participants have an up-to-date view of the leadership expected from their future selves.

However, the participants were also clear that a good leader challenged poor behaviour. This is not a leadership trait mentioned by Stoller et al (2004) or

Klaber et al. (2008). In focus group research one is exhorted not to forget the context in which the research takes place. Therefore, it may that the preceding questions regarding bringing the profession in disrepute, how to deal with the unprofessional behaviour of a colleague, etc. primed the participants to mention this aspect of leadership. However, in Good Medical Practice (2006a), the GMC states: "If you are responsible for leading a team, you must follow the guidance in *Management for doctors*" (p.22). Management for doctors (General Medical Council, 2006b) states that when leading a team you should:

> "monitor and regularly review the team's performance and take steps to correct deficiencies and improve quality" (p.9) and
>
> "deal openly and supportively with problems in the conduct, performance or health of team members through effective and well-publicised procedures" (p.9)

Therefore, while "challenging poor behaviour" was not mentioned as a leadership trait in some other focus groups, it does relate to some of the standards expected of leaders by the GMC.

Lastly, without being given any type of framework for discussing leadership qualities, the participants nevertheless referred to a number of the traits found in leadership frameworks such as that by Kouzes and Posner referred to in Stoller et al. (2004) (See Appendix 3-9). Klaber et al. (2008) do not provide any quotes to support the themes emerging from their focus groups on leadership. However, their themes are similar to the major topics of discussion in our groups, namely humility, confidence, expertise and the "ability to lead and work within teams". This may suggest that both sets of focus groups had a similar outlook on leadership.

Themes

The following themes were arrived at through an inductive process, by considering the answers to the questions posed, interaction between group participants and the greater context of medical students and professionalism. There are no rules for the number of themes to arise from focus group research.

Barbour (2007) suggested a maximum of 20, Hicks et al. (2001) derived 3, Paskins and Peile (2010) found 7. We would suggest it is the quality of the thematic analysis that is important rather than the number. The six themes discussed below were felt to be relevant and informative.

Theme 1: "Acting" versus "being" professional

All analysis requires some exploration of possibilities, provided that it is based on some observable, tangible findings. The participants frequently used the word "professional" both as a noun and as an adjective. However, when we consider the sentences in which the word is used, the participants seemed to use it interchangeably as an external manifestation (acting) and an internal state (being). There is no single incidence when a participant started to say "act/be professional", but then corrected him/herself to change the meaning.

It is possible that participants using the word "act" are referring to it in the sense of "do something" rather than "perform", however this is refuted by the number of instances in which participants refer to professionalism as a way of portrayal, how one is seen by others. There are at least two reasons why participants may have used "act" and "be" interchangeably.

The first is that, as with any skill or set of skills, including professionalism, one method of acquisition is through observation and emulation. Medical undergraduates can see the outward manifestations of professionalism, including dress, rapport with patients and colleagues, the ability to diagnose and manage illnesses, etc., and it is these outward manifestations which the undergraduates see as "professional". One can refer to Hilton and Slotnick's (2005) concept of proto-professionals, who observe and recognise professional behaviour without linking this back to the professional values from which the behaviour derives.

The second possibility is that medical students, as they develop their own professionalism, at times "act" professional without understanding why it is the correct thing to do. Because they are aware of this conflict, it may be that they

question whether everybody is merely "acting" professional or if some people actually "are" professional. A related matter is that medical students see much more unprofessional behaviour, in particular with regards to their peers, for a number of reasons including freedom from responsibilities and student culture. These same peers are then seen to be "acting" professional in the clinic, hospital or exams, contrary to what their fellow students know about them, a dilemma which has been raised by a number of authors (Ginsburg et al., 2004, Rees and Knight, 2007).

The act/be terminology may indicate that medical undergraduates would benefit from explicit description of the correlation between being professional and acting professional. The need for this guidance may be illustrated by the following quote: :

> "you're neither a professional nor in a professional setting [so I don't see why you should have to act professional" M3

Theme 2: The hidden curriculum

Although there was not a single use of the phrase "hidden curriculum" by the participants, the concept weaved its way through every focus group. As medical students progress through medical school, their moral reasoning deteriorates (Patenaude et al., 2003, Schillinger, 2006). A study from the USA (Satterwhite et al., 2000) reported that 24% of first year students thought that derogatory comments made about patients were sometimes or often appropriate. This percentage increased to 55% of fourth year students.

If we consider empathy, a construct which is related to moral reasoning, a systematic review showed that medical undergraduate empathy declines as they progress (Neumann et al., 2011). When considering the causes of this decline, the authors of the systematic review refer to aspects of the hidden curriculum such as:

1. Mistreatment by superiors or mentors
2. Vulnerability of medical students

3. Social support problems
4. High workload

The participants detailed a number of incidents of witnessing or being subjected to unprofessional behaviour by their "superiors". This result is supported by a survey of six medical schools which found that 98% of students witnessed unprofessional behaviour by their faculty (Feudtner et al., 1994). Some of the quotes also provided insight into the focus group students' vulnerability with regard to future employment prospects.

The fact that the hidden curriculum remains, if not the dominant force, then a major force in the medical student environment would suggest that a significant amount of work is still required to break this cycle. The consequences of witnessing or being the target of unprofessional behaviour are not just psychological (Rosenberg and Silver, 1984) but also include reduced task performance (Porath and Erez, 2007) and group dysfunction (Felps et al., 2006). These in turn lead to what Flin (2010) calls "a threat to patient safety and quality of care" (p.2480).

Theme 3: The rumour mill

One of the benefits about using focus groups is that it allows people to tell their stories. The "rumour mill" does not refer to these stories but rather to stories that the participants have heard, or vaguely recollected. It is natural, during the type of discussion that a focus group entails, for participants to mention things they've heard, that happened to someone else, on the "grapevine".

According to Kapferer (2013) rumours are usually spontaneous social productions, which arise when information is scarce. One of the problems with rumours is that they may not be based on facts, may be embellished and may have unwanted consequences. For example, Bucknall and Pynsent (2009) found that rumoured negative attitudes toward female orthopaedic surgeons was influencing the undergraduate teaching experience and career choice. In our focus groups, stories about "people getting away with things" and "students

being pulled up on things" do not portray a system where the students are aware of their responsibilities and the consequences of unprofessional behaviour. It may be that, to protect the individual, medical schools cannot provide specifics and that the rumours are an unfortunate but unpreventable consequence. Gerrity and Mahaffy (1998) refer to the adverse effects of rumour on effecting curricular change at medical schools; their advice to combat rumour is to communicate factual information. Similar advice is provided by Mennin and Krackov (1998) "addressing rumours and misinformation promptly was an essential communication process" (p.S62). Medical school faculties could consider sharing information about unprofessional behaviour and then providing an overview of decisions made nationally. Alternatively, it may be worth considering an open and transparent process, in which a consequence of unprofessional behaviour is appropriate remediation/punishment and exposure.

Theme 4: In the eye of the beholder

The two conversations presented in the Results are interesting for a number of reasons. They show that participants were willing to challenge one another; it would have been easy (and less onerous) for everybody to agree with the first speaker. The conversations also show the confusion, referred to by M11, caused by the lack of a "working definition of professionalism".

In addition, participants seem to confuse the idea that different patients expect different behaviours from the same doctor as an example of subjectivity. In fact, "Good Medical Practice" states:

> "To fulfil your role in the doctor-patient partnership you must: treat each patient as an individual" (p.15) (General Medical Council, 2006a)

It is therefore the professional doctor who changes his/her behaviour to that which is expected by the patient, in terms of manner of address, familiarity, jocularity, etc.

Lastly, the participants struggle with the fact that there are guidelines (from both the University and the GMC) regarding professional behaviour and the alleged subjectivity of professional behaviour. Some participants suggest that the guidelines set the boundaries of behaviour and the subjectivity then occurs within this:

> "…most people would say it's wrong to like to rob to rob other people and that that could be likened to you know in in hospitals certain behaviours are like kind of you don't have to mention them like people know that that's the way you should or shouldn't behave. But ehm there is there are more subjective areas which could be likened to like ehm in everyday life playing music loud at night kind of thing which isn't really morally wrong but a lot of people would might get annoyed. So that's kind of like different people that will have different you have different thresholds of what's what's eh what's seen as respectable and what what isn't kind of thing." M10

However, even in the example about loud music, there are guidelines such as bye-laws. Although it may be true that some people wouldn't mind the loud music while others would hate it, there are definite procedures for the latter to follow for the relevant authorities to make a decision regarding the negative impact on the complainants.

It is unclear why the subjective/objective confusion exists. It may be as a result of lack of familiarity with the topic or terminology. It may also be, with reflection on the hidden curriculum and the rumour mill, that participants see different responses to the same unprofessional behaviour and this reinforces the idea that what is or is not professional must be subjective. Lastly, there *is* a degree of subjectivity in the assessment of professionalism, as even the MPTS takes into account the specifics of each case in judging whether or not a given action was unprofessional.

Theme 5: The language of professionalism

In his text "Analyzing and reporting focus group results", Krueger (1997) counsels us to listen out for what is not said. In the discussions, there was

almost no mention of concepts such as beneficence, autonomy, non-maleficence, etc. This lack of discussion may have been a result of the questioning style or the relaxed, informal atmosphere. Another possibility is that the participants did not possess the necessary vocabulary.

In addition, the participants made numerous references to "GMC guidelines" but there was only a single mention of a GMC document "Duties of a Doctor". There is no such publication, and the reference is most likely to "Good Medical Practice" which, on the second page, has a headline "The duties of a doctor registered with the General Medical Council."

The linguistic relativity hypothesis suggests that the words we use shape our thinking and our worldview (Lucy, 1997). According to this hypothesis, not only does the absence of these key words suggest a superficial understanding of the foundations, structure and evolution of medical professionalism, but it also hinders reasoning, reflection and discourse. Although it is mere speculation, as the question was not asked of the focus groups, the author wonders if the discussion had centred around a pathology, such as chronic obstructive pulmonary disease (COPD), would the participants' vocabulary have been richer and more detailed?

Although one is able to map some of the discussions to the principles referred to in "Good Medical Practice" (Appendix 3-10), the lack of a framework or reference to GMC documentation suggests that the language of professionalism had yet to be adopted by the participants.

Theme 6: The gender of language

Wear and Kuczewski (2004) contend that: "…the rules, protocols, and expectations for physicians have always been developed *by* and *for* male physicians…" (p.3). The participants did not refer to this notion of professionalism as not being gender-neutral. This may be because it was not part of our questioning route. Alternatively, Wear and Kuczewski contend that professionalism as a male construct becomes more apparent when women are

mothers and wives and it may be that the majority of female participants were not married or mothers.

When analysing the use of he/she or gender-neutral talk, we did not look at quotes where the person referred to was obviously male or female. For example, if someone said: "My boss, she…" then that is not an example of the participant using gender-specific language. However if someone said: "A (non-specific) doctor has to think about his patients…" then that would be gender-specific. Focus group analysis is not a "counting game" where one compares the number of times something is said to decide what is most important. However, there were a number of instances where participants used the male pronoun when referring to a non-specific doctor. This may be because the male pronoun is quite often used as an alternative to the awkward he/she, but arguing against this is the frequent use of one/they in order to avoid he/she.

Three of the focus groups had a majority of one gender (FG 2: 4M/1F, FG3: 2M/5F, FG4: 4M/1F). One might hypothesise that the male-dominated ones were more likely to use the male pronoun but this was not the case. In addition, the only use of a female pronoun occurred in the male-dominated FG4.

There is the logical puzzle story in which a boy and his father are flying in a hot air balloon, which makes a crash landing. The boy is rushed to hospital where the surgeon, upon seeing the boy, says: "I'm sorry, I can't operate on him. He's my son." Many people will try and come up with an explanation including adoption or confused identity… The (simple) answer is that the surgeon is the boy's mother. Analysis of the language used by the participants may reflect this continued expectation that the "doctor" or the "surgeon" is a man.

The demographic change in Medicine, which means the majority of working doctors will be female, may lead to a reformulation of the "current dominant patriarchy" (Bleakley, 2013). Further studies assessing the gender of language of medical undergraduates may reveal such a change.

Generalizability

Much has been written in the focus group literature regarding the generalizability of focus group data. There are two main considerations regarding generalizability:

1. How accurately do the focus group data (verbal and non-verbal) reflect the true feelings, beliefs, attitudes, etc. of the participants?
2. How accurately do the focus group data (verbal and non-verbal) reflect the feelings, beliefs, attitudes, etc. of the population from which the sample was drawn.

There are differing opinions regarding the fundamental validity and reliability of focus group data (i.e. before taking transcription and analysis into account). How accurately does a focus group discussion reflect the beliefs of the participants? If a participant had a different pre-focus group day, e.g. argument with a colleague, difficult exam, then how different would their participation and responses be? Working within a constructionist framework, we should not pretend (as some do) that focus groups are naturally-occurring events, but rather "discussions occurring in a specific, controlled setting" Smithson (2000) (p.105). Sim (1998) claims that the participants are sharing a "public" account as opposed to a more private account they might share in an interview. He extends the claim by arguing that "Methodological considerations as to external validity therefore become redundant, as the whole enterprise of generalization is deemed to be misconceived at the outset" (p.350).

With regards to generalizability to the population, according to Bloor et. al (2001), focus groups are "not the authentic voice of the people" (p.15). Barbour (2005) suggests that "the goal of qualitative research is 'transferability' rather than statistical generalizability" (p.747) but then goes on to state that "theoretical generalizability" is a feasible goal. In the same paper, Barbour provides what she describes as a useful definition of 'theoretical generalizability' by Sim (1998):

> "Here, the data gained from a particular study provide theoretical insights which possess a sufficient degree of generality or universality to

> allow their projection to other contexts or situations which are comparable to that of the original study. The researcher recognises parallels, at a conceptual or theoretical level, between the case or situation studied and another case or situation, which may differ considerably in terms of the attributes or variables that it exhibits" (p.747)

Sim moderates this statement by saying that even this theoretical generalizability should be provisional. Sim also argues for different degrees of generalizability. He claims that it would not be unreasonable to postulate some degree of commonality between focus group members and others belonging to the same social category. Although this generalizability is not as rigorous as that expected of quantitative studies, it should not prevent us from forming hypotheses.

Limitations

The majority of the limitations of this study have already been referred to: the relative inexperience of the moderator, the inability to use the first focus group's audio data and the issues regarding generalizability of results. An additional limitation is the lack of respondent validation. Although this is often carried out in focus group research, Barbour (2005) says that it "is far from straightforward and its value will depend on the research. There can be ethical as well as practical problems and careful consideration should be given before providing written transcripts of group discussions" (p.748). We considered the confidential nature of the discussions, the personal disclosures and an inability to control the dissemination of the transcripts once released. As a result we decided not to email the transcripts and relied instead on the validation carried out at the time of the focus groups when a précis of the discussion was relayed and an opportunity for further clarification provided.

Conclusion

According to Marshall and Rossman (2010), analysis is sufficient "when critical categories are defined, relationships between them are established, and they are integrated into an elegant, credible interpretation" (p.209). We will address the two components of the study, professionalism and teamwork and leadership, separately.

Professionalism

Participants explored many aspects of professionalism, from the meaning and evolution of the term, to its relevance today and its impact on professionals at different stages of their careers.

The possibility that an increase in professionalism accompanies or perhaps is fuelled by an increase in responsibility deserves further attention. Medical students on clinical placements could be given minor and clearly defined responsibilities for patient care, with appropriate senior double-checking, such as clerking in patients, checking blood results, etc. Endowing the medical students with a role might integrate them into the clinical team, make them feel responsible for patient care and prepare them for professional practice (Evans and Roberts, 2006).

Although not referred to directly, in terms of the existence of a hidden curriculum which has a definite effect on medical student behaviour and attitudes, this focus group study adds further support to the literature. In addition, participants were unaware of the mechanisms for reporting unprofessional behaviour, did not display a vocabulary which suggests a working knowledge of professionalism and were only superficially knowledgeable of the guidelines which govern their behaviour and the behaviour of doctors. We hypothesise that this lack of knowledge sustains the hidden curriculum, as undergraduates are uncertain about what is or is not professional, and are afraid of the consequences of reporting unprofessional behaviour.

The above conclusion has two caveats. The first is that, if we accept the concept of proto-professionalism, as espoused by Hilton and Slotnick (2005), then 4th-year medical students should not be expected to be fully "professional". Hilton and Slotnick (2005) argue that professionalism is an acquired state rather than a trait, which "takes a number of years to attain" (p.59). This concept is supported by an interview study by Ginsburg and Lingard (2011) which found differences between pre-clerkship and clerkship students when considering professional standards. .Professional in this sense includes the idea that one knows the rules and regulations which govern the profession. One may therefore argue about the level of knowledge that 4th-year students should possess.

The second caveat is that the medical undergraduates' incomplete knowledge does not absolve the medical school or the hospitals in which the undergraduates are taught. Unprofessional behaviour directed towards medical students is not unique to Liverpool. McKegney (1989) refers to medical education as a "neglectful and abusive family system" (p.452), while Uhari et al. (1994) refer to the abuse of medical students and provide evidence of an international phenomenon. Until students are no longer exposed to unprofessional behaviour, especially when this originates from their "seniors", which is not subject to sanction, a determination to stamp out unprofessional behaviour in medical students is bound to fail. Focusing on students may, by some, be seen to be a worthwhile upstream exercise; when they become doctors their professional values, attitudes and behaviours will remain with them until they become the new "seniors". However, most research suggests that this is not the case; exposure to unpunished unprofessional behaviour is self-propagating. As Cooper (2002) writes:

> "We as leaders can't expect our students to succeed, while we model failure before them. . . . Actions speak louder than words. Professionalism is about walking the talk" (p.120)

We would not counsel the cessation of professionalism teaching; on the contrary, the teaching of professionalism (and its assessment) must form a greater part of the undergraduate curriculum. Students must learn and understand what professionalism is, so that they are equipped with the requisite knowledge to challenge (or at least identify) unprofessional behaviour in themselves and others. In addition, there must be a much more transparent process for reporting unprofessional behaviour, with appropriate safeguards for both reporter and reported. We would recommend that this process, as with the MPTS, is transparent regarding the adjudication process so that medical students no longer have to relay rumours to one another.

Teamwork and Leadership

The participants' notions regarding good and bad teamworkers and leaders seemed to be influenced by the milieu in which they work and study. Respect for one another, role clarity and a democratic leadership style reflect the importance that medical undergraduates' assign to these behaviours.

The concept of challenging poor behaviour was an interesting discovery. As referred to above, it is unclear why this concept was discussed in a number of the groups. It is possible that the preceding conversations regarding challenging unprofessional behaviour prompted the discussion. Although not referred to in the leadership framework provided in Appendix 9, the understanding that poor behaviour within a team must be challenged is not controversial. In fact, failure to effectively speak up about poor behaviour or observed mistakes has resulted in a number of well-publicised hospital deaths (Dyer, 2001, Dyer, 2004, Ferner, 2008). The attitudes and behaviours regarding teamwork and leadership discussed in these focus groups will inform the development of the assessment tool described in the next chapter.

Chapter 4: Development and evaluation of assessment tool

Introduction

This chapter will review the development and evaluation of the teamwork and leadership assessment tool. To gain broad acceptance, any tool must show that the resulting scores are valid and reliable. However, as Crossley et al. (2002) state: "All assessments must balance rigour (reliability and validity) against practicality (feasibility, cost and acceptability)" (p.803). Others add "educational impact" as an additional determinant of tool use outside the research setting (van der Vleuten and Schuwirth, 2005, Cook and Beckman, 2006). Though practicality and educational impact are not easily quantifiable (Norcini and McKinley, 2007) they must still be considered when determining assessment tool applicability outside of the research setting. The aim of this pilot study was to explore the practicality and possible educational impact of the tool, while also gathering data on reliability and validity.

Behavioural marker systems

Leadership and teamwork may be described as "non-technical skills". The first attempt at assessing non-technical skills in a medical setting as part of a behavioural marker system was carried out by Gaba et al. (1998) who modified an aviation checklist which included leadership as one of the assessed behaviours. The use of behavioural marker systems has since been widely adopted and adapted to rate non-technical skills in a number of healthcare settings (Gaba et al., 2001).

Simulation as a test-bed

There are a number of reasons for using a simulator setting in which to evaluate a tool. Brett-Fleegler et al. (2008) state: "Simulator-based rating systems have been used… with demonstration of good reliability and strong construct validity" (p.e598). In addition, the simulator setting can produce valid and reliable results (Devitt et al., 1998, Morgan and Cleave‐Hogg, 2000, Devitt et al., 2001, Murray et al., 2002). The life-sized mannequins can model critical events without the possibility of patient harm, the setting reflects clinical

practice and is suitable for evaluating technical and behavioural skills (Boulet et al., 2003, Ottestad et al., 2007, Lerner et al., 2009).

In addition, the ability to describe a set of actions does not correlate well with being able to perform those actions (Rethans et al., 1991). With respect to Miller's learning pyramid, high fidelity simulation prompts the participant to show that they are able to diagnose and manage the "patient" through good teamwork and leadership (Kyrkjebø et al., 2006). The participants "can demonstrate integration of prerequisite knowledge, skills, and affect in a realistic setting" (p.240) (Norcini and McKinley, 2007) and the realistic environment results in retention of learning through emotional involvement (Østergaard et al., 2004).

Simulation has been used to train healthcare personnel in teamwork and leadership (Helmreich, 2000, Grogan et al., 2004, Shapiro et al., 2004) and a Best Evidence in Medical Education (BEME) systematic review states that "high-fidelity medical simulations are educationally effective and simulation-based education complements medical education in patient care settings" (p.10) (Issenberg et al., 2005). In a study using simulation to teach the management of medical emergencies to undergraduates, 64% identified teamwork skills as a key learning point (Weller, 2004). Therefore using the same modality to evaluate teamwork and leadership seems reasonable (Srinivasan et al., 2006).

Lastly, simulation is used to summatively assess airline pilots on a yearly basis in high-stakes line operational evaluations (LOEs) (Baker and Dismukes, 2002). Although patients are not airplanes and doctors are not pilots, the acceptance of simulation as a mode of assessment by the aviation industry demonstrates the possibility, at least in theory, of using the same mode of assessment in healthcare.

Methods

The Methods section will cover pilot tool development and tool evaluation. In terms of nomenclature, the majority of behavioural marker tools follow a standard taxonomy. A **category** is an overarching term used to denote a desirable characteristic or trait, e.g. teamwork, leadership, situation awareness, decision making. Each category consists of a number of **elements**. Each element is an observable action. For example, in the NOTSS taxonomy (Yule et al., 2006), the category "decision making" consists of three elements:

1. "Considers options"
2. "Selects and communicates options" and
3. "Implements and reviews decisions"

The final term is **behaviour**. Each element may be performed either poorly or well, the behaviour describes the typical performance for a given rating. For example, in the element "Considers options" above, examples of good behaviours would be:

- Recognises and articulates problems
- Initiates balanced discussion of options, pros and cons with relevant team members
- Asks for opinion of other colleagues
- Discusses published guidelines

While examples of poor behaviours would be:

- No discussion of options
- Does not solicit views of other team members
- Ignores published guidelines

Therefore the typical taxonomy is Category:Element:Behaviour and these terms will be used in the remainder of this chapter.

Tool development

The elements used in the assessment tool were based on the results from the focus groups, the literature review, and additional sources (see below).

Focus groups

The focus groups were asked to discuss the characteristics of a good and bad teamworker and leader. The exemplary behaviours discussed by the focus groups were analysed and informed the decisions regarding which elements to include in the tool.

Literature review and additional sources

The literature review chapter analysed 23 articles in order to identify a tool that could be used to assess the leadership and teamwork skills of medical undergraduates. The tool characteristics and identified elements and behaviours of leadership and/or teamwork were extracted.

Existing assessment tools and literature which fell outwith the scope of the literature review were reviewed to provide additional leadership and teamwork elements and behaviours. This list of additional 23 papers was not meant to be exhaustive but instead included the expected teamwork and leadership elements and behaviours as detailed by the GMC, as well as other tools such as the Mayo High Performance Teamwork Scale (MHPTS) and the Observational Teamwork Assessment for Surgery (OTAS) tool.

In addition, the literature regarding the development and use of behavioural marker systems was referred to (Fletcher et al., 2000, Klampfer et al., 2001, Fletcher et al., 2003b, Thomas et al., 2004, Yule et al., 2006).

Elements

Grounded in a paradigm of critical theory and a constructionist epistemology, we approached the concepts of leadership and teamwork with the view that, not only are they social constructs, but that agreement on the "true" elements which comprise teamwork and leadership was neither feasible nor desirable. This does not mean that we embraced subjectivism, nor that any attempt to identify elements of leadership and teamwork would be worthless. Instead we appreciated that there were a multitude of elements and that the decision to include some, while excluding others, had to be defensible but could never be all-inclusive.

The first decision regarding the tool was the number of elements to be included under each category. The large number of elements and behaviours describing leadership and teamworking would need to be reduced to a number which would be feasible to evaluate in a simulated scenario. According to a seminal paper by Miller (1956) working memory capacity is 7 ± 2 items. We therefore decided that the maximum number of elements per category would be 5, which is in accordance with a number of other behavioural marking systems (Fletcher et al., 2003b, Yule et al., 2008).

The second decision consisted of the process of identifying the elements to be included in the tool. As detailed above, elements and behaviours considered to represent teamwork and leadership were collected from three sources:

1. The literature review texts evaluated in Chapter 2
2. Additional texts which did not fall within the scope of the literature review. These included further assessment tools and publications referring to teamwork and leadership such as "Medical students: professional values and fitness to practise" (General Medical Council and Medical Schools Council, 2009)
3. The focus group discussions detailed in Chapter 3

Every element identified by the three sources was established as a locus. Then every behaviour identified by the three sources was reviewed and either attributed to an existing locus or, if this was not possible, the behaviour became a new locus.
Using a method of triangulation, the loci from each source were compared to the loci from the other two sources in order to generate the final 10 elements.

Behaviours

For our simulation-based study, the behaviours were based on the possible performance of the participants within the scenario. Other behavioural marker systems already discussed in Chapter 2, e.g. ANTS (Fletcher et al., 2003b),

NOTSS (Yule et al., 2008) and NOTECHS (Mishra et al., 2008) were referred to when considering behaviour vocabulary.

Scoring system

We analysed the scoring systems of the above papers, in order to inform the development of the scoring system of our tool.

Assessment tool

Based on the information gathered during the tool development process, we divided the assessment tool into categories, elements and behaviours. For each category, we used a 5-point 5-item/element Likert scale, a 5-point global assessment score and a global assessment binary score to rate performance. The details are provided in the Results section below.

Tool evaluation

The tool was evaluated using a standardised simulated scenario. The assessors included the author and two specialist registrars involved in medical education.

Simulator

We used a METI® Human Patient Simulator (HPS) at the Cheshire and Merseyside Simulation Centre (CMSC). This high-fidelity mannequin has a complex software-driven physiology which is pharmacologically responsive and results in hardware-driven physical changes in the mannequin. One may administer drugs in real time and one may elicit breath and heart sounds, assess neurological function with eyelid and pupillary responses and feel for a full set of pulses. Measurement of a set of clinical parameters such as non-invasive blood pressure, pulse oximetry and ECG is also possible. The mannequin has a chest wall which expands and contracts with respiration and is able to model unilateral chest excursion which may be seen with a pneumothorax. The mannequin also allows for needle decompression of the thorax. The mannequin was controlled by an experienced operator using the METI® user-interface on an Apple™ computer. The operator and observer were situated behind one-way

glass which provided them with a view of the simulator suite. The operator also provided the voice of the mannequin via a speaker situated in the mannequin's head.

The various parameters referred to above are modifiable and a pre-determined sequence of events can be programmed in advance so as to provide a standardised change in physical and physiological status.

Scenario development

A number of pre-requisites would have to be met to ensure acceptability of the scenario:

a) The scenario would have to involve a presenting complaint which final year medical students would be expected to be familiar with
b) The scenario would have to be difficult enough to elicit relevant behaviours but not so difficult as to cause the participants to disengage at an early stage
c) The scenario would have to provide opportunities for the participants to display the behaviours linked to the elements in the assessment tool

Scenario design was undertaken with reference to published literature regarding the development of scenarios for evaluating behavioural skills (Bush et al., 2007) and by a process involving experienced members of the simulation centre faculty, experienced clinicians and three medical students who were not members of the test cohort.

The scenario development involved scripting of the mannequin's verbal responses to questioning, its baseline physiological status (Fig. 4-1) and response to predicted treatments,

Figure 4-1: Baseline settings for scenario

Patient: Tom Evans

Tom is 24 years old. He was brought in by ambulance from home. He is a known asthmatic who ran out of his inhalers. He has just arrived in A&E and is very short of breath. He does not have any notes or monitoring applied.
Airway: clear, speaking in short sentences
Breathing: widespread wheeze bilaterally, trachea central, Oxygen saturation: 90% on air
Circulation: Heart rate: 120, Blood pressure: 135/70. Normal heart sounds
Disability: Awake and alert
Exposure: Nil to add

as well as its deterioration over time. Scripting of the various extras was also required, in particular the confederates (see below in *Scenario running*) and the senior help who would allow the participants to display teamwork skills.

Scenario running

Every focus group participant was sent an email invitation to attend the one-hour simulation session.

At the beginning of the one-hour session, each medical student was welcomed, provided written consent, was briefed and given a routine introduction to the mannequin, its capabilities and limitations. This was followed by the scenario which lasted up to 15 minutes as this has been previously shown to be adequate time for accurate assessment of a candidate (Chambers et al., 2000).

The scenario was run according to the script by experienced mannequin operators and confederates. The role of the confederates, who take part in the simulation, is three-fold. The confederates ensure the scenario runs smoothly by clearing up misconceptions due to mannequin limitations or participant unfamiliarity. For example, if a participant states that they do not hear breath sounds when they are meant to be there, the confederates will correct them. The confederates will also provide information which cannot be gathered from the mannequin such as capillary refill time, skin temperature, colour, etc. The second part of the confederates' role in this scenario was to provide the participants with teamworkers so that their leadership skills might be assessed. Lastly, the confederates provide "standardised" team members with scripted responses. A lack of standardised team members has been mentioned as a limitation in other studies (Wright et al., 2009).

The scenario was identical for every medical student. The medical student was asked to perform within their expertise and examine and treat a patient with acute severe asthma in the A&E department. The mannequin then developed a tension pneumothorax requiring urgent treatment via needle decompression. The medical student was assisted by two confederates, playing the role of a

nurse and a healthcare assistant (HCA). Simulation centre staff performed both of these roles. Staff were allowed to prompt the participants according to standardised guidelines.

The first half of the scenario required the student to display leadership skills in dealing with a crisis and the second half was designed to allow them to show their team working skills. The transition from the first to the second half of the scenario occurred when a more senior doctor arrived on the scene. The senior doctor entered the scenario if one of the following conditions had been met:

1. The participant called for help
2. The participant had diagnosed the tension pneumothorax correctly and was proceeding to treat it by him/herself
3. The participant had failed to call for help by 8 minutes into the scenario

During this second part of the scenario we artificially created a conflict situation by scripting the senior help to make two potentially fatal mistakes: delaying needle decompression of the tension pneumothorax to await a chest x-ray, and decompressing the wrong side of the chest. We scripted this conflict for two reasons. The first was that our focus group research had shown that willingness to challenge poor performance was felt to be important. The second was that research in the assessment of professionalism, of which teamwork and leadership may be considered to be components, has suggested that the assessment should include a situation involving conflict (Ginsburg et al., 2000, Hafferty, 2006, Stern, 2006).

Every student took part individually in the same scenario and every student was asked not to disclose the particulars of the scenario to others.

Think-aloud, debrief and questionnaire

A think-aloud session followed the scenario, during which the medical student reviewed the tape of their performance and explained what their thoughts and feelings were during the scenario. The final 15 minutes consisted of a debrief led by one of the simulation faculty with advice regarding behaviour, non-

technical and clinical skills. This debrief used the assessment tool as a guide for feedback, allowing the medical students to see how they were being assessed. We felt that this debrief would provide a form of recompense to the medical students for investing their time.

The medical students then completed a questionnaire regarding the session, which allowed us to ask questions related to face validity and usability. (Appendix 4-1). The one-hour session was tested on a medical student in the same year-group who was not a member of the cohort and minor adjustments were made.

Feasibility

Feasibility was not formally evaluated but is addressed in the Discussion below.

Educational impact

Educational impact was evaluated using a post-assessment questionnaire of the participants in the simulated scenario.

Cost-effectiveness

Cost-effectiveness was not formally evaluated but is addressed in the Discussion below.

Acceptability

a) Participant/undergraduate

   Acceptability to the participants was evaluated using a post-assessment questionnaire of the participants in the simulated scenario.

b) Medical school, Regulator(s) & Public

   These aspects were not formally evaluated but are addressed in the Discussion below.

Validity

As discussed in Chapter 2, the classical view of validity classification was used (van der Vleuten, 2000):

- Construct

- The test is able to differentiate between different groups with known differences in ability. This evidence was not gathered.

- Content
  - Content validity was supported by carrying out the literature review (Chapter 2), as well as analysing additional assessment tools. Content experts in medical education and simulation-based medical education were also asked to provide feedback on the tool before it was used to evaluate the participants. Lastly, participants were asked if they thought that the scenario tested their teamwork and leadership skills.
- Criterion
  - The test is predictive of future performance or agrees with performance on a diffferent test carried out on the same day. This evidence was not gathered.

Reliability

Data were analysed using the SPSS® 16 (IBM SPSS, Armonk, New York, USA) software package.

a) Inter-rater reliability (IRR)

   Rater standardisation was performed using five videos and followed the phases of rater training described by Baker et al. (2001) of: information, demonstration, practice and feedback. The remaining videos were watched and rated independently by each rater. The raters had no knowledge of the clinical performance or exam performance of any of the students.

   The Intra-Class Correlation (ICC) is a ratio of the variance of interest over the sum of the variance of interest plus error, with values ranging from 0 (no agreement) to 1 (perfect agreement) (Shrout and Fleiss, 1979). There are several forms or models of the ICC. According to Nichols (1998), if there are an exact number of raters, who each rate all N persons, and the raters are not selected from a larger population of raters, then one should use a two-way mixed model to derive the ICC. The two-way model takes into account both inter- and intra-observer

variability (Cooper et al., 2010). In addition we were interested in absolute agreement, as opposed to consistency in scoring, we therefore used a two way mixed model with measures of absolute agreement (Nichols, 1998). We looked at average measures agreement. We used digital recordings to rate the participants (Swartz et al., 1997). As Brett-Fleegler et al. (2008) state, "reviewing videotaped performances to establish inter-rater reliability has ample precedent" (p.e601).

The ICC was calculated for each individual element, for the 5-point global rating score, the binary global rating score and for an average of the teamwork and leadership scores.

b) Internal consistency

Internal consistency was evaluated using Cronbach's alpha. The coefficient range is from 0 to 1, with values above 0.70 considered adequate (Sevdalis et al., 2008).

c) Test-retest reliability

This method of reliability testing was not carried out. Logistical challenges precluded the re-evaluation of the participants after a time delay, or the re-evaluation of the videos by the same raters.

Results

Tool development

Designation of elements

The elements and behaviours from 3 sources were used in a process of triangulation in order to inform the elements to be used in the final assessment tool. The three sources were:

1. The leadership and teamwork elements and behaviours from each of the 23 papers reviewed in chapter 2 (Appendix 4-2) (Source 1)
2. The leadership and teamwork elements and behaviours from existing assessment tools or research articles which did not fall within the scope of our literature review analysis (Appendix 4-3) (Source 2)
3. The leadership and teamwork elements and behaviours from the focus group discussions (Table 4-1) (Source 3)

Table 4-1: Focus group leadership and teamwork behaviours

Categories	Elements	Behaviours
Leadership	Situation awareness	Looks over everything
		Gathers information
	Communicates well with team	Respects team members
		Diplomatic
		Good communicator
	Role allocation and workload distribution	Distributes tasks
		Shares the load
		Controls team members
		Knows role and role of others in team
	Goal declaration and updating	Makes casting vote
		Decisive
		Inspires and motivates
		Shares common goal
	Information gathering	Gathers information

	Maintenance of standards	Maintains standards
		Quality control
Teamworking	Feedback	Voices his/her opinion
		Communicates well
		Good communicator
	Task acceptance and completion	Accepts responsibility for carrying out role
		Works as part of the team
		Willing to lead
	Team-member support	Shares the load
		Respects team members
	Challenging poor performance	Willing to confront team members who are underperforming

The process of identifying elements and behaviours referred to in the Methods section above resulted in 22, 18 and 12 loci from the literature review sources, the non-literature review sources and the focus groups, respectively. (See Appendix 4-4)

Triangulation resulted in the 10 final elements used in our tool. In this process some elements were not included in the final tool, for example "Dress and appearance", because it was felt that they would not be good markers for teamwork or leadership. Other elements were combined, for example the two elements "Role allocation" and "Workload distribution" resulted in a final element of "Allocates roles/tasks to appropriate team-members and ensures workload is shared". Appendix 4-5 details the results of the triangulation process.

The final list of elements was:

- Teamwork
    - Accepts and completes tasks
    - Provides appropriate feedback to team leader and team-members
    - Adopts leadership role if necessary
    - Supports other team-members
    - Challenges leader if appropriate
- Leadership
    - Listens to team members and responds appropriately
    - Allocates roles/tasks to appropriate team-members and ensures workload is shared
    - Declares goal and how to achieve it, changing this if necessary as new information is collected
    - Maintains situational awareness or ensures SA is maintained by another if leader distracted
    - Solicits opinions from team-members

Behaviours

Every element was provided with a behaviour for every Likert point. For example, for "Teamwork: Accepts and completes tasks" the behaviours were:

- Very good: Provides history, carries out competent ABCD.
- Good: Provides history, carries out ABCD but misses out components.
- Acceptable: Provides history, carries out ABCD but misses out A or B or C or D
- Poor: Does not provide history or does not carry out ABCD
- Very poor: Refuses or ignores request to accept task when directly asked

Scoring systems

The scoring systems used by the above-mentioned papers are provided in Appendices 4-2 and 4-3. The majority of the assessment tools used Likert scales, ranging from 4 to 24 items and 4 to 9 points. A small number of tools used

Yes/No or Yes/No/Borderline checklists. Lastly, some assessment tools included a “not applicable” or “not observed” point.

Assessment tool

The two categories we used were teamwork and leadership. Each category consisted of 5 elements and each element listed examples of behaviours which would be considered good, poor, etc. The assessment tool is provided in Appendix 4-6.

Tool evaluation

Scenario development

The simulation scenario consisted of a standardised script in which a young adult male with a known history of asthma was admitted to hospital with an acute severe asthma attack. The patient deteriorated over a given time-frame with physiological parameters which would be better or worse depending on participant actions. The patient went on to develop a life-threatening tension pneumothorax which required needle decompression. Once this procedure had been carried out the patient’s status improved and, after the participants had the opportunity to discuss further management, the scenario was brought to a halt.

Scenario running

The scenario was run based on the script referred to above. Monitoring such as automatic intermittent BP measurements via a cuff, electrocardiograph and pulse oximetry was available if requested. Continuous results were displayed on a monitor. All equipment that might be expected to be available in an Emergency Department was available for use. In addition, clinical paraphernalia such as drug chart, observation chart and medical notes were created and made available. The role of the confederates were modelled on an emergency department nurse and a health care assistant.

Talk-aloud, debrief and questionnaire

Details of the talk-aloud component of the sessions are provided in Chapter 5. There are no results from the debrief, except for a question regarding its usefulness which is detailed below in "Educational impact". Results from the questionnaire are detailed in the appropriate sections below.

Feasibility

The assessment in its current form takes 30 minutes in the simulator-naïve participant, as mannequin and environment familiarisation takes approximately 15 minutes. There is a minimum of 2 members of staff, one controls the mannequin and the other acts as a confederate in the simulation room. Our set-up involved 4 members of staff, one controlled the mannequin, one observed (and debriefed the participant) and two members of staff acted as confederates.

The assessment tool involves selecting a point on a total of 12 5-point Likert items and 2 2-point scales while observing the 15-minute video performance.

Educational impact

The debrief used the assessment tool as a guide for feedback. In response to the question "Did you find the debrief where we discussed your personal teamwork and leadership useful?" 100% (29/29) of the participants answered in the affirmative. 72% (21/29) of the participants provided a free text answer which is provided in Table 4-7.

Table 4-7: Free text responses to "Did you find the debrief useful?"

Candidate	Response:
1	It is great at my time to get personal feedback, and a brilliant way to learn about your abilities
2	Good to hear that it is a common experience that students are reluctant and wary of taking on a leading role
3	I think simulation experience is fantastic because it can be

	very realistic and gets the adrenaline pumping. The debrief allows me the opportunity for constructive feedback and it is very important to use your team so you can keep your attention focused on the patient
5	I felt as if there may have been more negative aspects we could have discussed but maybe I'm being paranoid rather than you holding back.
6	Helped me to identify my strengths and weaknesses
8	Partly useful as the positive aspects of my performance were highlighted and appreciated but I was also looking for constructive criticism - which did come but after probing!
9	It allowed me to reflect on the good and bad things which I did during the scenario. It also gave me a chance to see my weaknesses and how I can improve them
10	Gave points to improve and positives
13	I think discussion of errors and strongpoints meant the exercise was more worthwhile for my own personal learning
14	This will help me to build upon my professional team work and leadership skills and hopefully help me to react more appropriately in similar situations in the future
15	Filled me with some confidence that I did the right things as I felt that I hadn't performed very well
17	Yes, in this case, it made me more confident for future
18	It made me realise that sometimes I need to communicate my rationale and decisions to allow others to have an input
20	Very useful for personal development and future practice
21	Will definitely be helpful in future real life scenarios
22	It was nice to have some positive feedback
23	For my own personal learning
25	Helped to identify areas for improvement
26	Like getting feedback

28	Made me think about how I work in a team specifically following instructions
29	Really good to know where to improve. Great practise and good way to consolidate what you learnt

In addition a number of free text answers to the question "Is there anything about the hour that you think we should change? Could we improve the experience in any way?" suggested a positive educational impact (Table 4-8).

Table 4-8: Free text responses suggesting positive educational impact

Candidate	Response:
3	I think it was excellent. Obviously it is a study but I felt I was tested and challenged appropriately and it is only in these sorts of situations that you learn how you might react in real life.
9	The simulation is very realistic and it is very useful training scenario
20	I was very pleased with the experience. It was educational for me.
21	Very helpful experience for me to see how I work in this type of scenario

Acceptability

100% (29/29) of the participants answered the question "Was the introduction to the sim centre and the mannequin adequate?" in the affirmative. 55% (16/29) of the participants provided an additional free text response to the above question. These are provided in Table 4-3 below:

Table 4-3: Free text responses to "Was the introduction to the sim centre and the mannequin adequate?"

Candidate	Response:
2	Given enough information, but not too much - directed as to the elements I'd be needed to use
3	A lot of detailed information given about equipment for all eventualities but I knew exactly what was available to hand
6	It is a complicated piece of machinery and so it's really good to get a detailed explanation
9	The limitations of the mannequin and the roles of the assistants were explained well
10	Knew where everything was and how to use things I needed
14	I was informed of where to find all the equipment I might need and about how the mannequin worked, which was really helpful
15	Very thorough. Got me nervous by showing me equipment I have never used before
17	Very well done. Mark took me through everything very thoroughly. I felt nervous at first but the team put me at ease
20	It was very useful to look around the room beforehand and see where everything was and what was available
21	Useful to know that I could ask for senior help at any point, and have equipment explained
23	I was fully aware of all relevant equipment
24	Good run through of equipment and mannequin
25	I was already familiar with the layout but this was useful to clarify!
26	Been on MEDSIM before
28	Good to refresh knowledge about where stuff was even though we've been before

29	Good directions around the centre and on how the mannequin works. Good that we were allowed to listen to the chest first

100% of the participants answered the question "Do you think that the scenario and assessment was fair and acceptable to you as a medical student?" in the affirmative. 79% (23/29) provided an additional free text response which is detailed in Table 4-9:

Table 4-9: Free text responses to "Do you think the scenario and assessment was fair and acceptable?"

Candidate	Response:
1	I feel it was really useful, both for my own skills and clinical knowledge
2	I was worried that it may be a trauma situation, of which I have little experience, but the use of a common and core emergency scenario that all final year medical students should be familiar with, and have knowledge of the management of was fair
3	I think being alone in a medical emergency is a nightmare thought for any student. It was clearly an asthma attack, although I'm a little disappointed not to have diagnosed the tension pneumothorax faster but as medical students we should definitely know about them
4	It was quite a difficult standard, and a sharp learning curve
6	It detailed a common and important condition
7	Increased my awareness of my own abilities, e.g. ask for a registrar before performing a pleuritic tap
9	This is a typical scenario that would be seen in A&E and one which as a junior doctor I would be expected to manage
12	I would say that it was challenging as I have not

	encountered it in real life
13	The scenario of an asthma attack was appropriate to our level
14	I feel that I should have in theory have been able to cope with the situation I was faced with. Despite the fact that I struggled, I do feel this was a fair assessment
15	Covers areas we should be familiar with. Asthma and pneumothorax. Especially after finals!!
17	Yes. Good/brilliant practice for future. We don't always get this scenario as a medic in hospital/or we are limited in terms of what we can do.
18	As a final year student we need to be comfortable with management of common emergencies and know when to seek senior help
20	It was very relevant to a 5th year and someone who is about to start foundation training
21	Found that it was an appropriate scenario for our knowledge level - even if I can't remember how to read ABGs!
22	Probably quite reflective of the scenarios I will be facing as an F1 in a few months time
24	Yes as 5th year medical student about to be a F1 it is important. Slightly in the deep end.
25	This is a scenario that a medical student will never be in, but an F1 doctor could easily be in
26	Scenario was level we should know
27	We are expected to be able to deal with emergency especially common ones like tension pneumothorax
28	Common clinical scenario which we should know how to treat. More difficult part was challenging the senior but this does happen and it's good to practice in non-threatening environment

29	Realistic situation which I will be expected to deal with in practice

The question "How realistic was the whole scenario?" may also reflect the acceptability of the assessment. This visual analogue score ranged from 0 = absolutely unrealistic to 100 = as real as real life. The mean (±SD) was 73 (±12).

In addition a number of free text answers to the question "Is there anything about the hour that you think we should change? Could we improve the experience in any way?" suggested acceptability (Table 4-10).

Table 4-10: Free text responses suggesting acceptability

Candidate	Response:
1	It was brilliant actually, especially seeing yourself in a clinical setting and it is so rare to be able to do so
6	It was well structured.
14	It was a bit scary, but everybody was extremely friendly and helpful. I don't think there were any areas which could be improved.
17	Everything was very well set
18	It was excellent
20	I was very pleased with the experience.
24	More experiences throughout medical school. Compulsory sessions.
26	Very good.
28	Really useful and good to think about methods of challenging people who are more senior. Well organised. Very informal - good!

Validity

Content

a. Representation: The theoretical constructs of leadership and teamwork have been explored by a number of different sources. These sources postulated various elements which would be representative of the constructs. The assessment tool is based on a triangulation of these sources and could therefore be considered to reflect a broad perspective of the two constructs.
b. Face: Face validity is supported by a review of the assessment tool by content matter experts and experienced assessorrs of teamwork and leadership: Dr Helen O'Sullivan, Dr Arpan Guha, Mr Peter Leadbetter, Mr Ray Fewtrell, Ms Jayne Garner, Dr Simon Mercer, Mr James Goulding and Mr Neal Jones.

   Face validity is also supported by the 100% (29/29) of participants who answered the question "Do you think that the scenario tested your leadership and teamworking skills?" in the affirmative. 69% (20/29) provided additional free text responses which are detailed in Table 4-2:

Table 4-2: Free text responses to "Do you think that the scenario tested your leadership and teamworking skills?"

Candidate	Response:
2	I'm used to, and comfortable in a junior role and being told what to do. I still feel too inexperienced to adopt a more commanding authoritative role. The simulation definitely highlighted that.
3	It was realistic in the sense that I was an F1 and had a nurse and HCA available. I tried to use them both to share workload but I didn't know exactly how much a HCA could do.
4	It tested the leadership skills very well due to the range of situations it put you in.
6	It forced me into a situation where I had to make decisions

	and couldn't just rely on other people to make them
9	I was able to be involved as part of the team and co-ordinated the roles of the nurse and the HCA. It also involved challenging those in leadership which can be important in healthcare
10	Being placed as a leader at the start and then have a senior enter later on
13	A good mixture of both leadership and teamworking in a realistic environment
14	This situation tested my ability to be a leader and also to follow instructions - although artificial I feel this is a similar situation to which I will be placed as a junior doctor
15	Was very much like real life with the whole team involved
17	Definitely tested leadership skills, in fact I have learnt not to blindly trust a senior!
18	I had to ask for investigations and assistance. Once the senior doc arrived I had to challenge his view on the tension pneumothorax
21	Found it very helpful to have nurse and HCA and to be able to use their expertise in the situation
22	I was put into a scenario where I was initially the leader and had to co-ordinate the team of 3. Later, a senior arrived and this tested my ability to communicate and resolve disagreement/conflict
23	Having to politely deal with seniors, who were interrupting an urgent procedure was something I had not considered before
24	Subtly tested, good test of delegation, communication, leadership and teamworking
25	It made me realise how to improve such as determining skills at the start and the communication skills regarding conflict with colleagues

26	Always good to get tested in safe environment
27	A bit more challenging than what we did in MEDSIM as here you feel you need to be doing things and tell people what to do. You can't just step back and think properly
28	Leadership - Yes because initially you are alone and making decisions. Teamworking - Yes because you then become the junior
29	I played the leader in the scenario and had to work with various other members of the unit. Some members (i.e. anaesthetist) were difficult to deal with

Reliability

a) Inter-rater reliability (IRR)

For both ICC and Cronbach's alpha, values greater than 0.6, 0.7, 0.8 and 0.9 are classed as minimally acceptable, respectable, very good and excellent, respectively (DeVellis, 1991).

Each scenario was scored independently by the three observers. The results for the ICC of the 24 scenarios are shown in Table 4-4 below.

Table 4-4: Intra-class correlation (ICC)

Element	ICC
TW1	0.74
TW2	0.73
TW3	0.73
TW4	1
TW5	0.88
TWG1	0.81
TWG2	0.71
L1	0.62
L2	0.71

L3	0.73
L4	0.80
L5	0.75
LG1	0.72
LG2	0.78

The results for the averages of scores, i.e. the sum of the scores of the elements in each category divided by the number of elements scored are shown below (Table 4-5).

Table 4-5: Intra-class correlation (average measures)

Element	ICC (average measures)
TWAverage	0.85
LAverage	0.77

b) Internal consistency

Internal consistency was evaluated using Cronbach's alpha (Table 4-6).

Table 4-6: Cronbach's alpha

Category	Cronbach's alpha coefficient
Teamwork	0.85
Leadership	0.81

Discussion

Tool development

In the development of any assessment tool there are a number of opposing tensions (van der Vleuten and Schuwirth, 2005). The tool should be short enough (fewer elements) to be used within a given time-frame but long enough (more elements) to fully encompass the category being considered. The tool must also be simple enough (fewer points within a given Likert item) to allow the rater to make a decision, but complex enough (more points) to allow the rater to differentiate between performances. The implementation of the findings from the literature review and focus groups are discussed below.

Literature review and additional sources

The majority of papers which used a taxonomy followed the structure of: Category:Element:Behaviour. For example, "Situation awareness: Gathering information: Reduces level of monitoring because of distractions" (Fletcher et al., 2003b). The exception is Moorthy et al. (2005) who follow the taxonomy: Behaviour:Element, e.g. "Preoperative preparation: Introduction to team members". It was felt that the "Category:Element:Behaviour" taxonomy was the more logical and this is what our tool would be based on.

Although many of the assessment tools specified whether they were referring to leadership or teamwork elements or behaviours, some did not. As explained in the Results, in these cases the author made a decision on the classification. It should be noted that the default was for an element or behaviour to be classified as "teamwork". Primarily because attributes such as "trustworthy", "good communication" and "shares information or resources" are not unique to leadership. In this view, leaders are team workers with additional elements characteristic of leadership.

Even a cursory examination of Appendices 4-2 and 4-3 makes it evident that there are a plethora of elements and behaviours which respective authors suggest describe teamwork and leadership. The need for a distillation of these

myriad elements is self-evident. The final tool was developed with information from the focus groups.

Focus group discussions

Focus group analysis does not support giving greater weight to behaviours or characteristics which are referred to more frequently. The results of the focus group discussions were therefore presented without a frequency table. In order to increase the acceptability of the tool to those being evaluated, the results of the focus group discussions were incorporated into the distillation process referred to above in order to derive the final elements discussed below.

Elements

The assessment tool had two categories: Teamwork and Leadership. The derivation of the elements from the vast number of existing elements would necessarily involve a degree of subjectivity. The elements may also be considered to cover more than one behaviour, e.g. "accepts AND completes tasks" "declares goal AND how to achieve it", "allocates tasks AND ensures workload is shared". Tool development, in particular discussion with content experts, showed the need for linking of behaviours. This is supported by van der Vleuten and Schuwirth (2005) who argue against the "atomisation" of skills. In addition, the correctness of decisions regarding the number of elements, the use of a Likert scale and the number of Likert points, although based on the available evidence regarding assessment tools, would not be fully established until the evaluation phase.

Behaviours

We used a descriptive behavioural marker scale as opposed to a numerical scale as it has been suggested that assessment tools should be providing observers and participants with the standards expected of them, as opposed to a number (Academy of Medical Royal Colleges, 2009). The descriptive behaviours or "anchor statements" also improve the reliability of a rating scale (Thistlethwaite and Spencer, 2008) and reduce personal bias in interpreting performance (Kim et al., 2006).

Scoring systems

The majority of the existing tools use Likert scales. There are a number of benefits of Likert scales when evaluating behaviour:

- The use of a number of points allows the observer to evaluate the candidate within a range of possible behaviours, e.g. from "very poor" to "very good", as opposed to a binary "yes/no" evaluation
- Using a number of points allows one to evaluate a possible improvement in performance over time, e.g. from "acceptable" to "good" to "very good" and to provide formative feedback to a candidate
- A greater number of points allows the observer to be more discriminating e.g. rating a candidate as "poor" or "very poor"

Likert scales have some drawbacks, which must be taken into account when using them within an assessment tool:

- Observers must be standardised to agree on distinctions between points, i.e. what distinguishes "good" from "very good" performance
- As the number of points increases, the greater the theoretical discriminatory power, however in practice there is a pay-off between number of points and IRR
- Although Likert scales have a rank order (e.g. from 0 = "very poor" to 4 = "very good") the intervals between the values are not equal, i.e. the difference in performance between "very poor" to "poor" is not necessarily the same as from "poor" to "acceptable" (Jamieson, 2004)
- Raters have a tendency to avoid the extremes of the Likert points and cluster around the middle

The tool uses an odd number of Likert items. Although there is debate regarding the use of an even or odd number of items (Croasmun and Ostrom, 2011) the author felt that the ability to choose a mid-point item would allow raters to evaluate a performance as average, adequate or acceptable rather than "above average" or "below average". This position is supported by the large number of

tools which use a mid-point and by Garland (1991) who states: "the explicit offer of a mid-point is largely one of individual researcher preference."

The tool uses 5-point Likert items. The debate regarding the ideal number of points has been referred to above. Symonds (1924) argued that the ideal number was 7, while Jamieson (2004) informs us that the usual number is 5. The majority of behavioural assessment tools use a 5-point item scale, as it seems this provides the correct balance between discriminatory power and ease of scoring. More important than the number of points is the need to calculate the internal consistency of the items. This is discussed below in the evaluation of the tool.

The tool provided an "unable to assess" option. It seemed self-evident that if an element was not seen the raters should be able to record this. Raters had to be informed that this was option was to be selected if they were unable to assess a given element, rather than if an element was poorly performed. For example, there is a difference between the participant not listening to team members (which would be rated as "poor" or "very poor") and not being able to assess this element, for example because there were no team members. The distinction between these two ratings has proved difficult in other behavioural marker systems (Fletcher et al., 2003a).

In addition, we used two global rating scores for each category, one 5-point Likert and one binary (acceptable/unacceptable). The global rating scores allowed raters to provide another evaluation of the participant without focusing on specific elements. This approach is widely supported (Cox, 1990, Cohen et al., 1991, Cunnington et al., 1996, Morgan et al., 2001a, Govaerts et al., 2002). A number of studies have found that global rating scores are at least comparable with checklist scores in terms of reliability and validity (Keynan et al., 1987, Cohen et al., 1991, Regehr et al., 1998, Swartz et al., 1999). Regehr et al. (1998) state: "Global rating scales scored by experts showed higher inter-station reliability, better construct validity, and better concurrent validity than did checklists." In addition, the binary score allowed raters to focus their judgment

to evaluate the participant's performance as either "acceptable" or "unacceptable". This binary score could also be used in summative assessments as a "pass/fail" evaluation. One disadvantage of global ratings is that they may mask deficits in particular skills (Brett-Fleegler et al., 2008) and therefore the use of a checklist-type Likert scale and a global rating should provide the best of both assessments.

Lastly, we did not use a weighted scoring system. A weighted scoring system might consider that, for example, supporting other team members is considered twice as important (and therefore collects twice as many "points") as adopting a leadership role. Our tool development did not encompass a Delphi-type assessment of the importance of the given elements and we considered that, for our purposes, a decision about the level of performance of a given element was sufficient.

Tool evaluation

Simulator

The mannequin used was a "high-fidelity" model because this was the only mannequin in use in the simulation centre at the time. There is however no reason why a "medium-fidelity" model could not be used instead. This would result in significant cost savings if the tool were to be used outwith a simulation centre with "in situ" simulation.

Scenario development

The scenario development was uncontroversial, the simulation centre staff had a significant amount of experience in developing scenarios which were appropriate to the level of the participants. Participants were able to display the sought-after behaviours and the degree of difficulty was not such that participants disengaged at an early stage.

Scenario running

Simulation centre staff had run a large number of simulation courses with group sizes of up to 16 people. Running scenarios for one person at a time was therefore relatively straightforward. We were able to stagger the start-times to allow overlap and therefore run up to 6 1-hour sessions in 4 hours.

Talk-aloud, debrief and questionnaire

The talk-aloud is discussed in Chapter 5. The debrief and answers to the questionnaire are discussed in relevant sections below.

Feasibility

One of the benefits of an assessment tool which can be used to rate videotaped performances is that the raters do not need to be present when the assessment is administered. This increases the feasibility of the tool as raters can watch the videotapes when it is convenient for them.

In terms of using a simulator, Norcini and McKinley (2007) state that simulators "are very expensive, they require considerable space and staff support, and the development of cases and scoring requires significant expert input" (p.243). However, the simulation centre already runs courses that every final year medical student at the University of Liverpool must attend. It would be feasible to run 15 minute scenarios with 15 minute debriefs, which would allow one to assess 8 medical students in 4 hours. In aviation simulations, the scenario is used to evaluate both the captain and first officer (Baker and Dismukes, 2002). If our scenario were run with one medical student leading and one being the team worker then this would further increase the utility of every scenario. In addition, the scenarios can be run with the minimum of 2 staff members. For high stakes assessment the scenario could be marked from video-recordings which allow the observer to replay instances and make sure no observation was missed.

Performance-based assessment studies suggest that the number of scenarios is more important than the number of raters and that reliability and validity are improved by having a number of different scenarios rather than our single 15

minute scenario (Boulet et al., 2003, Rall and Gaba, 2005). The number of scenarios required is unclear, and would be based on the intended use of the assessment (e.g. formative, summative).

The following adaptations could be made to increase the number of scenarios:

- Delay the feedback/debrief until the videotapes have been evaluated
- Provide feedback/debrief via a proforma in electronic format
- Increase the number of participants in each scenario, e.g. 1 leader and 3 team workers and rate every participant

One could also use the scenarios to rate technical skills such as basic airway manoeuvres or assessment of the cardio-respiratory system. This would improve the efficiency of the scenarios and may reduce the need for Objective Structured Clinical Examinations (OSCEs).

When Cruess et al. (2006) carried out semi-structured interviews to ascertain the limitations of P-MEX, they found that the major limitation is time: time to train raters, time to observe, time to record and time to feedback. Further research will need to focus on the time and number of scenarios needed to obtain a stable result for a given student.

Educational impact

Educational impact means "the effect of the assessment, positive or otherwise, on students' learning and development" (p.5) (General Medical Council, 2011). In terms of the scenarios, simulation scenarios have been shown to help students learn, using the assessment as a formative experience (Issenberg et al., 2005). In addition to the experiential benefits, the use of simulation and subsequent debriefing provided study candidates with immediate feedback on their performance. Feedback has been shown to improve non-technical skills acquisition (Savoldelli et al., 2006) and supports a General Medical Council (2011) recommendation:

> "Good feedback will be effective in improving learning and performance" (p.18)

However, medical undergraduates feel that feedback is lacking both in quality and in quantity (Urquhart et al., 2014). 100% of the candidates in our study found the debrief useful. Although this was not an in-depth response to the debrief, a positive educational impact may be postulated from the free text responses such as:

- "Really good to know where to improve. Great practise and good way to consolidate what you learnt"
- "Helped me identify my strengths and weaknesses"
- "It allowed me to reflect on the good and bad things which I did during the scenario. It also gave me a chance to see my weaknesses and how I can improve them"

In addition, Urquhart et al. (2014) interview and focus group study with medical undergraduates found that a positive feedback experience resulted from feedback which was constructive, specific, based on direct observation, balanced, and respectful. It is likely that many of these conditions were met during the debrief.

Debriefing also promotes the type of reflective behaviour which encourages and sustains professionalism (Myerson, 1998). Lastly, in "Assessment in undergraduate medical education: Advice supplementary to Tomorrow's Doctors" the General Medical Council (2011) states:

> "Simulated environments can also provide effective assessment opportunities. As *Tomorrow's Doctors* (2009) states at paragraph 100: 'Medical schools should take advantage of new technologies, including simulation, to deliver teaching'; and at paragraph 102: 'Opportunities should also be provided for students to learn with other health and social care students, including the use of simulated training environments with audio-visual recording and behavioural debriefing'. Simulation can be

> appropriate to assess both technical and non-technical skills." (paragraph 86)

In terms of the assessment tool, Ian Hart is quoting as saying that students "learn not what you expect, but what you inspect" (p.41) (ten Cate and de Haes, 2000). This quote is backed up by experimental evidence (Newble and Jaeger, 1983, Frederiksen, 1984, Broadfoot, 1996). Students may therefore place greater emphasis on displaying the behaviours detailed in the assessment tool (Schuwirth and van der Vleuten, 2010). It is hoped that such a change would be a positive one.

Cost-effectiveness

In terms of the assessment tool, costs were incurred to fund this research, however there are no additional costs to use this tool. In terms of the assessment costs, many of the arguments have been addressed in the Feasibility discussion above.

In terms of computer-based case simulations, the reliability per unit of testing time is less for the simulation than for multiple-choice questions (Clauser et al., 2002). However, it could be argued that there is no other evaluation which allows the undergraduate to demonstrate a holistic assessment, diagnosis and management of an unwell patient in such a short time-frame. In addition this set of competencies, skills and behaviours is then assessed by observers who do not work with the participant and therefore do not have any of the typical biases which arise in that situation. Using a videotape to analyse performance also has cost benefits, as raters can carry out evaluations outwith the simulation centre and at a later date (Devitt et al., 1998, Georgiou and Lockey, 2010).

Costs may also be reduced by employing raters who are not medically qualified personnel; a number of studies have shown that their reliability can be as good as medical personnel, as long as they have had adequate training (Martin et al., 1996, Fraind et al., 2002, Slagle et al., 2002). In fact, some studies suggest that non-medical observers are better at assessing inter-personal factors, such as

teamwork and leadership (Schaefer et al., 1994, Schaefer et al., 1995, Carthey, 2003). The use of non-medical observers may have an impact on acceptability, but perhaps would be accepted in low stakes, formative assessments. The use of peers to rate clinical skills is also gaining increased acceptance (Perera et al., 2010, Moineau et al., 2011). Basehore et al. (2014) found that peers were able to accurately rate "complex clinical skills" in an OSCE. However, there is a paucity of evidence in the use of peers to rate non-technical skills such as teamwork and leadership, particularly in a realistic, simulated environment. The financial benefits of using peers are obvious, but additional research in this area is required.

Although not an improvement in cost-effectiveness, the removal of certain aspects of medical selection, such as entrance interviews, which have very low validity evidence (Salvatori, 2001, Eva et al., 2004, McManus et al., 2005), would allow redistribution of funds to tools with higher validity evidence. The argument regarding cost-effectiveness of simulation is supported by Hofmann (2009) who concludes: "simulation can be effective and efficient in the education of hi-tech health care."

Acceptability

Undergraduates

All of the undergraduates felt that the introduction to the centre and to the mannequin was acceptable. In addition, the scenario scored highly for realism. These results are supported by a focus group study of final year medical students, who comment on the emotional realism of simulation-based teaching and also on how participants found simulation to be an ideal way of developing team working and leadership skills (Paskins and Peile, 2010). The realism of high-fidelity simulation is supported by Gaba (2004): "experience shows that participants in immersive simulations easily suspend disbelief and speak and act much as they do in their real jobs" (p.i2).

100% of the students found that the assessment tool was fair and acceptable. Although the numbers are small, this contrasts strongly with a previous study by Duffield and Spencer (2002) which showed that, across a range of assessments, the maximum percentage of students who considered a given assessment to be fair was 78%.

In addition, the students found that our tool was a valid test of their teamwork and leadership skills, with 100% agreeing with the statement: "Do you think that the scenario tested your leadership and team working skills?".

Raters

The raters consisted of the author and two anaesthetic specialist trainees with an interest in simulation and medical education. Although not formally assessed, by the nature of the development of the assessment tool, all three raters found the tool to be acceptable.

Regulators

A number of researchers have shown that workplace-based assessments are subject to bias (Streiner, 1995, Paisley et al., 2005) or performed poorly (Day et al., 1990, Noel et al., 1992, Holmboe, 2004b) with very poor inter-rater reliabilities (Streiner (1985) referenced in van der Vleuten et al. (1991)) . In addition, a tool which can be used to assess performance in a simulated environment means that a number of factors can be standardised. This includes predictable deterioration in "patient" physiology and scripted responses by assistants. The simulated scenarios can therefore be repeated for large cohorts which should increase the acceptability to regulators.

Much of the current assessment of the "shows how" competence level in Miller's pyramid, particularly in the final examination, is carried out using the OSCE. However, as van der Vleuten (2000) argues, the OSCE does not reflect clinical reality as it often relies on assessing a single skill, e.g. examination of the knee, in a restricted time period. The OSCE can therefore be forcing the candidate to act at a lower level of competence. In addition, the correlation between the

OSCE and written tests are high, which suggests that little is achieved by carrying out both (Vleuten et al., 1989). The benefit of the simulated exercise is that, although it takes 15 minutes, it reflects a true scenario in which a patient deteriorates, and needs treatment, in real time.

In their Best Evidence in Medical Education (BEME) review, Issenberg and Scalese (2007) argue that the ability to provide a range of task difficulty levels, appropriate to the learner, is one of the key features of simulation-based learning. We can envisage the use of this assessment tool longitudinally throughout a medical student's undergraduate course. Leadership and team working skills of the first year medical student may not be best demonstrated by a scenario requiring extensive knowledge. However, even the response to a simulated cardiac arrest may allow us to assess these skills in any medical undergraduate. Do they call for help? Do they lead a team or adopt a team working role?

Public

A formal investigation of the acceptability of the tool or the assessment to the public was not carried out. However, one may speculate that an assessment which allows the participants to demonstrate a range of teamwork and leadership elements, in real time and in an environment where no patient will be harmed, would be considered acceptable to the public.

Validity

Content validity refers to "the representativeness of the test blueprint achievement domain" (p.2168) (Kim et al., 2006) or the extent to which a measure represents all facets of a concept. Content validity for our tool would mean that the elements listed under teamwork and leadership encompass these two concepts in this setting. The elements were based on a literature review of existing tools, input from medical undergraduates, and had been reviewed by educationalists and simulation experts, as recommended by Slocumb and Cole (1991).

Some components of leadership and teamwork might be assessed in a paper-based exercise. In particular, the theoretical knowledge underpinning leadership styles or an understanding of cognitive biases might be explored using single-best answer or multiple choice questions. However, Boulet et al. (2003) have shown that there may be significant disparities between knowing what to do and doing it. Therefore, in order to assess teamwork and leadership skills and behaviours, the assessment tool should be designed for use in a context which is as realistic as possible, such as a simulated scenario or a clinical context (van Mook et al., 2009a). Therefore, the use of a tool which is based on a comprehensive literature review, examined by content experts and based on carrying out an evaluation in a realistic setting, supports a claim to content validity. Lastly, the participants themselves unanimously agreed that the scenario tested their leadership and teamworking skills.

In terms of criterion validity, we did not gather data from other assessments that the participants undertook, e.g. written examinations, OSCEs, end-of-placement reports. Evidence from elsewhere suggests that there is poor correlation between simulation-based assessment and other assessments (Morgan et al., 2001b). As has been argued in this chapter, the simulation-based assessment may be testing different levels of performance than written tests or tests examining a narrow skillset. It is therefore possible that the results of the simulation assessment would differ from the results obtained in other assessments, but that this would not mean that the simulation assessment was inaccurate.

In terms of construct validity, Kim et al. (2006) state that this component of validity evidence would be supported by showing that there is a difference in levels of training. However, as referred to in the Conclusion below, it is possible that simulation assessment will not distinguish between different levels of undergraduate training as it is assessing applied knowledge (shows how) which most medical undergraduates, irrespective of years of training, struggle with. In addition, Sevdalis et al. (2008) warn against the traditional validation approach

of demonstrating "differences across different levels of expertise" with non-technical skills, because it is unclear if these skills naturally increase with time in training.

Reliability

Rater standardisation was performed using five videos, which has been shown to be a sufficient number in a similar set-up (Moorthy et al., 2006).

a) Inter-rater reliability (IRR)

   The IRR for the individual elements varied from minimally acceptable (0.62, L1) to excellent (1, TW4); both of these scores deserve further attention.

   TW4 was "Supports other team-members". The custom-made scenario did not include any built-in occurrences where this element could be demonstrated. This was done on purpose to see whether the raters would reliably and appropriately mark "not observed" for this element. The raters appropriately marked "not observed" for 23 scenarios and in the only scenario where a candidate carried out an unscripted behaviour which suggested support for a team-member all 3 raters marked the candidate as excellent. This perfect agreement between raters results in an ICC of 1. Future uses of this assessment tool should include a scenario where this element is included in the scenario design.

   L1 was "Listens to team-members and responds appropriately". The wording for the behaviour was: "Takes in information from team-members and…"

   - "…only occasionally acknowledges receipt/acts on information" (Acceptable) or
   - "…mostly acknowledges receipt/acts on information" (Good) or
   - "… shows understanding by repeating salient points frequently and always acting on information." (Very Good)

For L1, none of the candidates scored less than “Acceptable” and the majority of candidates scored “Good”. However there was a lack of agreement between raters scoring “Acceptable” and “Good”. In order to improve the ICC score for this element we must consider rewording of the behaviour to remove the ambiguity between “only occasionally” and “mostly”.

IRR was also acceptable when asking the raters to rate the candidate’s team working or leadership as a pass/fail decision (acceptable or unacceptable) with respectable agreement (0.71 for team working and 0.78 for leadership) between raters.

In their paper, Graham et al. (2010) showed improved ICC using the averages of scores, i.e. the sum of the scores of the elements in each category divided by the number of elements scored. This calculation also removes the problem of the “not observed” category in the ICC calculations. We also found an improved ICC for this averaging of scores, with teamwork (0.85) and leadership (0.77) scores. It seems that this would be a worthwhile score to provide to candidates along with their individual scores.

b) Internal consistency

This refers to the statistical or psychometric data (Kim et al., 2006) e.g. items which are meant to be scoring the same (or similar) variable are more closely correlated than items scoring different variables. High internal consistency suggests that the elements are measuring the same characteristic, e.g. teamwork or leadership. For example, if a person scores poorly on one teamwork item he or she should perform poorly on other teamwork items.

Internal consistency of the scoring system, i.e. whether the elements represent the entire scale and are consistent with each other, was

evaluated using Cronbach's alpha. Construct validity for our assessment tool, in terms of internal consistency, was high. Cronbach's alpha coefficient values >0.7 are typically considered adequate (Sevdalis et al., 2008) and our values for both teamwork and leadership were >0.8. This suggests that the elements are measuring the same characteristic. In addition, our values of 0.81 and 0.85 suggest "commonality but not duplication" (Fletcher et al., 2003b).

Limitations

Psychometrics

Due to logistical and time constraints we were only able to run one scenario per undergraduate. As a result some psychometric tests, specifically those requiring more than one test, such as test/re-test and inter-cases reliability, and criterion validity, were not achieved. According to Schuwirth and Van der Vleuten (2003) "inter-rater reliability is a relatively small source of error compared with inter-case variability" (p.69). In addition, the literature suggests that a number of simulated scenarios are necessary to enhance validity and reliability (Boulet et al., 2003); Epstein (2007) suggests a minimum of 10 scenarios when using simulated patients.

Due to the small numbers we did not analyse subgroups of students, e.g. those who did not have English as a first language. It has been shown that poor communication and performance may be due to a lack of competence in the given language (Cushing et al., 2014). This may therefore account for some performance issues in our study.

The elements used to describe teamwork and leadership showed high internal consistency, i.e. they seem to be measuring the same construct. However, unlike Cooper et al. (2010), we did not carry out a formal assessment of the relevance of each element with content experts. It is unclear how productive such an exercise would have been, given the number of teamwork and leadership elements and behaviours that have been described. Cooper et al. (2010) used six

content experts to carry out a rating of the relevance of their teamwork elements but it is likely that a much larger number would be required to provide a universally accepted ranking of elements. Due to these considerations we asked content experts to provide feedback on the assessment tool in an informal manner as described in the Methods section.

Rater training

The raters were standardised using a small number of videos; it is possible that more extensive standardisation may have resulted in greater IRR. Further evaluation of the tool is required in terms of rater training and we did not carry out usability testing. The raters in this research were all senior anaesthetic trainees with a background in medical education and simulation. It is unclear how long other raters would require to undergo standardisation and the authors of other rating tools have suggested that rater training may take up to 2 days (Klampfer et al., 2001, Yule et al., 2009), although this was for a wider assessment of behaviours (Flin2010). Some authors suggest that the use of frame of reference (FOR) training for raters, with standardisation against vignettes displaying a spectrum of behaviours, improves rating accuracy (Noonan and Sulsky, 2001, Roch and O'Sullivan, 2003).

Rater standardisation was carried out using IRR training, i.e. after every one of the first five videos the raters compared scores, discussed discrepancies and decided on how to score similar behaviours in future scenarios. Although this is accepted practice, Goldsmith and Johnson (2002) argue that IRR training may lead to reliable but inaccurate scores, as the focus is on the raters rather than on the performance. They argue instead for gold-standard training, where raters' scores are compared to a gold-standard score set by experts. This may be an area worthy of further exploration with respect to this assessment tool.

We made no attempts to analyse the differences between raters in terms of observation accuracy and rating accuracy (Baker and Dismukes, 2002). Poor IRR may be a result of certain raters not seeing the observed good or bad behaviour (observation accuracy) or the result of seeing the behaviour but then

scoring it inappropriately (rating accuracy) (Carthey, 2003). Further research in this areas is required in order to inform decisions regarding where rater training is most required.

The development of the tool was not as extensive as that carried out by Fletcher et al. (2004) for their ANTS system, which involved a literature review, an examination of existing marker systems, cognitive task analysis interviews, an iterative development process involving workshops and cross-checking in theatre. This was due to constraints in terms of time, personnel and finances. However, the ANTS rating system, despite its extensive development, has met with a number of practical problems such as inter-rater reliability and rater training (Graham et al., 2010). In the end, one must decide if our rating tool is acceptable to the various stakeholders and usable outside the research setting.

Conclusion

We conclude by considering the place for this assessment tool in the constellation of current assessments, aspects requiring further research and the need to balance the costs of this type of assessment with the benefits of evaluating realistic performance.

DiMatteo and DiNicola (1981) call for multiple, subjective assessments from different sources in the evaluation of the performance. This need for triangulation of assessments is widely supported (Thistlethwaite and Spencer, 2008, van Mook et al., 2009c, van der Vleuten et al., 2010) and is echoed by Hawkins et al. (2009) who state that "multi-dimensional constructs require the application of multi-modal assessment approaches" (p.352). This assessment tool may therefore form a component of professionalism evaluation which can be supported by other assessment methods; if we wish to examine all levels of Miller's pyramid then we must use assessments which are appropriate to each. Additionally this assessment tool will provide participants with the feedback they require to become better team workers and leaders. This dovetails with the argument for a shift in the rationale of assessment, from "assessment of learning" to "assessment for learning" (van der Vleuten, 2012, Dannefer, 2013).

Further research is required to see how performance in terms of teamwork and leadership is associated with academic performance. In addition, further research is required on tool psychometrics. However, the need for psychometric rigour above all else as a measure of the value of an assessment has been challenged, with the arrival of a "post-psychometric era" (Eva and Hodges, 2012, Hodges, 2013, Southgate and van der Vleuten, 2014). Southgate and van der Vleuten (2014) argue that "Judgement by knowledgeable people is imperative for assessing complex performances... and (many) subjective judgements may lead to defensible high-stakes decisions." Devitt et al. (2001) have shown that it is possible to use a simulation-based evaluation to discriminate between undergraduates and postgraduates, and between different postgraduate experience levels. However, Young et al. (2007), although finding a difference in scores difference between undergraduates and postgraduates, found no

difference in performance between undergraduate year groups in high-risk scenarios. The authors suggest that the undergraduates, in lecture-based and PBL-based courses, are not being prepared for the practical clinical challenges ahead. Boulet et al. (2003) obtained similar results in their simulation-based study.

As discussed, we did not correlate simulation assessment data with other data, e.g. test scores or pass rates. From the preceding paragraph, it would seem unclear how useful this data would have been, as the likelihood is that the simulation evaluates a different aspect of the participants' performance than other tests. In particular, the "response to a crisis" nature of the simulation scenario is not evaluated at any other stage of the undergraduate curriculum. This means that, if Hilton and Slotnick (2005) are correct in their speculation regarding the existence of a spectrum of professionalism (proto-professionalism), then it is possible that this tool may not be able to identify different stages of undergraduate development

In addition, the number of scenarios which would provide a minimum range of content to allow for variability in performance is unclear. In OSCEs, for example, the amount of time is measured in hours (Petrusa (2002) as referenced in (van der Vleuten and Schuwirth, 2005). Boulet et al. (2003) found only moderate correlation when using six scenarios, although they used a checklist and did not supplement this with global rating scores. There is also some suggestion that inter-case variability is a much greater source of error than inter-rater reliability (Schuwirth and Van der Vleuten, 2003).

Further research is also required to determine ideal scenario length. We chose 15 minutes for a number of reasons. Logistically 15 minutes meant we could run one simulation every hour (when introduction, talk-aloud, debrief and questionnaire-completion were included.) In addition, the simulation centre scenarios are usually of a 15-minute duration, as this seems to provide the right amount of balance between allowing participants to perform and providing material for the debrief. It is unclear if scenarios could be shorter, Schuwirth

and Van der Vleuten (2003) warn against making the scenarios too short and therefore less realistic.

There are a number of caveats which must be considered when using this tool. The first is that the elements, although reviewed and approved by a group of experienced educators, were developed without external input. Further development of this tool using a Delphi process would provide academic rigour. The second caveat is that the tool was developed for use in a simulated environment where the scenarios are pre-planned to provide undergraduates with the opportunity to show good or bad behaviours. It is therefore unclear how this tool would translate into a clinical environment. The third caveat concerns the findings regarding reliability and validity. Both reliability and validity are properties of the scores and not the tool. As Cook and Beckman (2006) state: "The same instrument, used in a different setting or with different subjects, can demonstrate wide variation in reliability" (p.e13) and, by correlation, validity. The third caveat concerns two of the elements: "Adopts leadership role if necessary" and "Challenges leader if appropriate". These were felt to be important attributes of a good teamworker. However, outside of the simulation environment, they may occur infrequently. Should this tool be used in the clinical environment, one should reflect on the need for elements which are important but rare; this is one of the drawbacks of a behavioural marker system (Klampfer et al., 2001). In addition, as detailed in the Discussion section, the L1 element's behaviours may need re-phrasing to create a greater distinction between "mostly" and "only occasionally". One of the benefits of using frequency-based descriptive behaviours is the removal of the difficulty of what Gaba describes as "aggregating a single rating for a behaviour that fluctuated over time." However, the corollary is that distinctions between the frequencies must be clear.

One could debate whether high-fidelity simulation directs participants to demonstrate performance (shows how) or action (does) (Schuwirth and Van der Vleuten, 2003). If the participants act as they would in real life then the latter is arguably the case. Brown and Doshi (2006) argue that we must move

away from one-off assessments to workplace-based assessments. (WPBAs) They argue that one-off assessments, such as the OSCE, "do not assess other attributes necessary for a person to perform consistently well as a doctor, for example team-working skills." However, this assessment tool suggests that one can use high-fidelity simulation to demonstrate those skills without the many drawbacks associated with WPBAs (Holmboe, 2004a).

The adoption of this assessment tool may have a wider educational impact than merely what has been discussed with respect to those being assessed. ten Cate and de Haes (2000) argue that, as their numbers grow, assessors will begin to reflect on their own leadership and teamwork skills. This would have the beneficial effect of emphasising the importance of these skills to assessors who will be more senior healthcare professionals. In addition, if faculty are involved in assessing the behaviours they will gain some insight into the strengths and weaknesses of the students and, by extrapolation, some indication of where further teaching is best focused (Martin et al., 1996).

According to a 1998 survey of 24 UK medical schools, it is not concerns regarding validity or reliability that are the major obstacle to a new assessment, but rather "lack of staff time and of resources" (Fowell et al., 2000). Unfortunately, major changes in medical education, the professionalism expected of students and doctors, as well as the loss of the "apprentice" model of graduate training mean that methods of assessment must change. Students want to know that they are being assessed fairly and equitably, with summative decisions based on actual performance rather than rater reliability issues. The public wants to be able to trust junior doctors, particularly given the findings of an increase in mortality when new doctors start in August (Jen et al., 2009). Regulators want to be assured that medical schools are producing doctors who are "fit for purpose" from day one. We would argue that assessing how medical students actually perform in a realistic environment requiring an integration of knowledge, skills and behaviours should be a major form of assessment in the medical school curriculum. As Swartz et al. (1999) state: "performance

assessment… requires the observation of performance". We believe that the staff time and resources to achieve this must be allocated.

To conclude, this tool for assessing teamwork and leadership skills can be used in a manner which is feasible, acceptable and cost-effective, while further work is required on element-formation, reliability and validity.

Chapter 5: Challenge the leader

Introduction

One of the elements in our teamwork and leadership assessment tool was: "Challenges leader if appropriate". As discussed in Chapters 3 and 4, this element, although perhaps only occasionally required in the clinical setting, is an important part of effective communication and teamwork (Mahlmeister, 2005). In addition, as referred to in Chapter 1, failure to challenge poor leadership has been implicated in a number of small and large incidents and accidents, both within and outwith Medicine.

In order to allow the participants in our research project to be evaluated on the element "Challenges leader if appropriate", we scripted a scenario in which they would be provided with the opportunity to challenge poor leadership. This gave us the ability to funnel the medical undergraduates into two situations where they would be forced to either challenge a poor decision or ignore it. These discordances would occur in a realistic environment, in a crisis, with real-time decisions required of the participants about whether or not to speak up. This provided us with the opportunity to video-tape the encounter and then ask the participants to talk about their reasoning while reviewing their own performance. We could therefore be provided with a glimpse into the thought processes of final-year medical undergraduates who delayed in challenging or failed to challenge a poor decision from a more senior member of healthcare staff.

The concept of exploring post-performance reasoning is not entirely new. Sheehan et al. (1987) carried out a research project in which the participants underwent an "ethics" scenario with a simulated patient and then "provided information about their attitudes and intentions" in a post-scenario interview. Using a post-evaluation oral or written assessment to explore reasoning during an ethics OSCE is also suggested by Lynch et al. (2004). In addition, while discussing the assessment of medical student behaviour, Ginsburg et al. (2004) stated that they "should look beyond observable behaviors to include the

reasoning behind them, in order to develop a more accurate assessment of a student's developing professionalism" (p.S1). Lastly, Rees and Knight (2007) refer to a Reflective Judgment Interview (RJI) which can be used to assess students' reasoning strategies. However they state that the RJI should be used to evaluate the students' ability to reason through ill-structured problems rather than their own behaviour after an event.

This investigation, which is a component of the larger study on the assessment of teamwork and leadership, should provide us with some insight into the behaviour and reasoning of medical undergraduates when placed into a situation which requires them to speak up against an authority gradient.

Methods

The recruitment of the candidates for this study has been detailed in Chapter 3. The running order of the simulation sessions, whereof this analysis forms a part, has been detailed in Chapter 4, however it is worth revisiting the salient points.

The leader had to make two erroneous decisions in order to allow for the possibility of a challenge. This meant that the scenario had to be forced into two checkpoints before each erroneous decision. The flowcharts in Appendix 5-1A and Appendix 5-1B detail the running of the scenarios based on the action or inaction of the participant.

The "leader" entered the scenario as a result of three possible events:

1. The participant called for help ("leader" entered at 8 minutes)
2. The participant had diagnosed the tension pneumothorax correctly and was proceeding to treat it by him/herself ("leader" entered before 8 minutes)
3. The participant had failed to call for help by 8 minutes into the scenario and "leader" entered as a passing-by senior doctor

In order for the challenge to be possible, the leader had to make two poor decisions (Table 5-1).

Table 5-1: Two erroneous decisions

Erroneous decision 1	Delaying needle decompression of a tension pneumothorax in order to await a chest X-ray which will take 15-20 minutes to be carried out. Vital signs are: BP 70/50, HR 130, SpO2: 75
Erroneous decision 2	Decompressing the wrong hemi-thorax. The "leader" plans to decompress the wrong hemi-thorax despite the absence of breath sounds, hyper-resonance to percussion and deviation of the trachea away from the other (correct)

	hemi-thorax.

In order for these poor decisions to manifest, a number of conditions had to be met. These conditions included:

- the need for the leader to prevent a very able (or over-confident) candidate from attempting to treat the tension pneumothorax without calling for help. In such a case, the leader therefore entered the scenario prematurely in order to create the challenge
- the need for the participant to be sure what the diagnosis was. The leader therefore asked the participant to re-assess the patient and provided prompts, as required, until the candidate diagnosed (and said the words) "tension pneumothorax"
- the need for the participant to be sure that the incorrect side of the thorax was going to be decompressed. The leader therefore created a nonsensical theory that in a tension pneumothorax the trachea always deviated towards the side needing decompression and that he/she would therefore decompress the (incorrect) right-hand side.

The dialogue surrounding the leadership challenges was analysed, as was the post-scenario review of their thought processes by the participants. Lastly, the questionnaire completed at the end of the activity included questions relating to the talk-aloud technique and the results from these are detailed below.

Pian-Smith analysis of scenario dialogue

The videos of all 29 participants were reviewed and the dialogue between the participants and the "leader" was transcribed. This dialogue was coded according to a scoring system developed by Pian-Smith et al. (2009) (Table 5-2).

Table 5-2: Pian-Smith Scoring system

Type	Score	Example
Say nothing	1	
Say something oblique,	2	"Really?"

obtuse		
Advocate or inquire	3	"I'm concerned about this platelet count" OR "Don't you think this platelet count is too low?"
Advocate or inquire repeatedly; with initiation of discussion	4	"Can we talk about this platelet count?" AND/OR "I'm uncomfortable with these platelets" AND/OR "What do you think?"
Use crisp advocacy-inquiry	5	"I'm wondering about risks of doing this when there's a low platelet count. How do you decide how to proceed?"

Analysis of rationale for behaviour using "think aloud" technique

All 29 participants undertook a review of their performance immediately after the scenario. While watching the digital video-recording of the scenario, they were asked to tell the researcher what they were thinking using a retrospective "think-aloud" technique (Fonteyn et al., 1993). The participant was encouraged to maintain a continuous monologue about their cognitive processes. As recommended by Lewis and Rieman (1993) gentle prompting from the researcher, such as "Tell me what you are thinking" and "Keep talking", was used if there was a prolonged pause or if the participant deviated on to talking about other issues such as their actions or performance. These reflections were recorded onto a digital audiotape and the passages relating specifically to the instances of leadership challenge were transcribed and analysed for ostensible rationales for failure to challenge.

The rationale(s) for failure to challenge provided by the participant were categorised according to a list of possible causes adopted from Pian-Smith et al. (2009) (Table 5-3)

Table 5-3: Reasons for failing to challenge (Pian-Smith et al., 2009)

Perceived barriers to action	**Additional barriers when challenging a teacher or mentor**
Assumed hierarchy Fear of embarrassment of self or others Concern over being misjudged Fear of being wrong Fear of retribution Jeopardizing an on-going relationship Natural avoidance of conflict Concern for reputation	Respect for the teacher/student relationship Violation of a special trust High value placed on experience Concern over being negatively evaluated

Post-scenario questionnaire

The questionnaire included two questions specific to the think-aloud episode.

1) Was the think-aloud technique acceptable in terms of ease of performance?
   Yes/No
   Please elaborate
2) Were you able to remember why you did/said things during the think aloud technique or did you feel like you had to make up things?
   Able to remember/Had to make up things/Bit of both
   Please elaborate

In addition, some responses to the question "Is there anything about the hour that you think we should change? Could we improve the experience in any way?" referred to the think aloud technique.

Results

Due to a hardware fault the final 11 recordings of the think-aloud technique were not recovered from the digital audiotape. The data are therefore presented for both the overall 29 participants or the first 18 participants dependent on data availability.

Pian-Smith analysis of scenario dialogue

29 participants were offered 2 challenges each for a total of 58 possible challenges. These could be placed into 3 categories (Table 5-4).

Table 5-4: Classification of challenges

Category	N	%
Challenged without prompting	37	64
Challenged after prompting	12	21
Never challenged	9	15

The Pian-Smith scores of these 58 challenges were classified as follows (Table 5-5)

Table 5-5: Pian-Smith classification of challenges

Score	**N**	**%**	**Example**
1 (Say nothing)	10	17	
2 (Say something oblique)	2	3	But his sats have already gone down…
3 (Advocate/inquire)	9	16	Have we got hyper-resonance here?
4 (Advocate/inquire repeatedly)	25	43	His trachea's shifted to this side. It might be a

			tension it might be worth decompressing before the chest x-ray.
5 (Crisp advocacy & inquiry)	12	20	Are we able to decompress him before the chest x-ray or do we need to confirm it with the chest x-ray? But ehm normally normally I've been taught to decompress it as soon as we can.

If we consider only those participants for whom the think aloud recordings are available, the responses of the first 18 participants are classified as in Table 5-6.

Table 5-6: Classification of first 36 challenges

Category	N	%
Challenged without prompting	20	56
Challenged after prompting	10	28
Never challenged	6	16

The Pian-Smith codings of these 36 challenges were classified as follows (Table 5-7)

Table 5-7: Pian-Smith classification of first 36 challenges

Score	N	%	Example
1 (Say nothing)	7	19	
2 (Say something oblique)	2	6	But his sats have already gone down.
3 (Advocate/inquire)	8	22	Shouldn't we do something about the pneumothorax?
4 (Advocate/inquire repeatedly)	16	44	I'm just wondering if the trachea is deviated. I wonder if we would possibly… Do you agree?
5 (Crisp advocacy & inquiry)	3	8	There's less breath sounds. Which way is the trachea deviated? No, I disagree.

Analysis of rationale for behaviour using "think aloud" technique

In 10 instances the participants only challenged the erroneous decision after prompting from the leader and/or nursing staff.

In 7 instances (Table 5-7) the participants did not verbally challenge the "leader". In one of these instances the participant did not verbally challenge but instead picked up a needle and went to carry out a needle decompression. This explains the discrepancy between the number of people who "say nothing" in Table 5-7 (7) and those who "never challenged" in Table 5-6 (6). Therefore, in 6 instances participants did not challenge despite repeated prompting from the leader and/or the nursing staff. In these cases the nursing staff had to challenge the leader's erroneous decision.

These 16 instances of delayed challenge or lack of challenge were analysed. In 15 instances the participant was aware that the senior was making an erroneous decision and in 1 instance it is unclear (Appendix 5-2). In 3 instances within the talk-aloud technique the participants did not provide a rationale for not challenging. The remaining 13 instances were classified as follows (Table 5-8). The numbers add up to more than 13 because in some instances participants provided more than one rationale.

Table 5-6: Rationales for not challenging or delaying challenge

Reason	Number (Percent)	Example
Assumed hierarchy	10 (77%)	*"…because the anaesthetist had arrived and I felt that we'd kind of transferred responsibility to him…"*
High value placed on experience	6 (46%)	*"…he's obviously had experience of this in the past…"*
Fear of being wrong	2 (15%)	*"… I had to understand it before we proceeded cos we needed to get this right."*
Fear of embarrassment of self	1 (8%)	*"Didn't want to say something that was gonna to be completely ridiculous."*

Post-scenario questionnaire

Responses to Question 5: "Was the think-aloud technique acceptable in terms of ease of performance? Yes/No (Please elaborate)" are detailed in Table 5-9.

Table 5-9: Responses to Question 5

Reply	Number (Percent)	Elaboration
Yes	26 (90%)	*"I actually prefer thinking aloud, in OSCEs and what nots I prefer to do this so it wasn't a problem. I also think it helps the team*

		know what I'm doing" *"I felt at ease talking through my thought process"* *"Personally I prefer verbalising my thoughts aloud in situations like this"*
No	3 (10%)	*"It can difficult to view yourself immediately after the performance"* *"Didn't really have much to say. Struggled to fill the silence."* *"It is difficult to know what I was thinking at the time"*

Responses to Question 6: "Were you able to remember why you did/said things during the think aloud technique or did you feel like you had to make things up? Able to remember/Had to make things up/Bit of both (Please elaborate)" are detailed in Table 5-10.

Table 5-10: Responses to Question 6.

Reply	**Number (Percent)**	**Elaboration**
Able to remember	25 (86%)	*"It was strange to begin with saying exactly how I felt but I got really into it and relaxed to say exactly what was running through my mind"* *"I did remember as throughout the scenario I try to signpost what I'm doing"* *"Never made anything up though often couldn't remember what I was thinking"* *"As it was done immediately afterwards the thoughts were fresh in my mind"*

Had to make things up	0	
Bit of both	4 (14%)	*"I was able to remember some of the reasons why I acted certain ways during the scenario. Although it can be difficult to describe spare of the moment decision"* *"At some points I don't think I was thinking anything at all, other than "help!!""* *"Slightly difficult. Prompted by visuals and memory however you cannot help but notice things that went wrong."*

Responses to Question 8: "Is there anything about the hour that you think we should change? Could we improve the experience in any way?" which referred to the think-aloud technique are detailed in Table 5-11

Table 5-11: Responses to Question 8

Free text answer
"Everything was very well set. Just give more information/instructions on 'aloud' technique so that the student knows what to say."
"I would have preferred to have watched myself first through in silence, between giving feedback on my performance"

Discussion

Pian-Smith analysis of scenario dialogue

The study by Pian-Smith et al. (2009) carried out pre- and post-intervention analysis of simulated scenarios which provided anaesthesia trainees with three opportunities to challenge erroneous decisions by other healthcare workers. The intervention was a discussion and teaching session on using the advocacy-inquiry method to challenge decisions. Our data are more in keeping with the post-intervention group of their study. When our participants did challenge the erroneous decision, the most frequently used technique was repeated advocacy and/or inquiry.

These results agree more with a simulator-based study by St Pierre et al. (2012) which looked at the willingness of residents and nursing staff to challenge deliberate errors committed by attending physicians. The authors modified the Pian-Smith model, amalgamating crisp advocacy-inquiry with repeated advocacy-inquiry. In this study, the participants were more likely to use crisp/repeated advocacy inquiry (40%) than an oblique statement (35%).

It is unclear why our participants, who were not formally trained in advocacy-inquiry, used this as frequently as the post-debrief group in Pian-Smith et al.'s study. One possibility is that Pian-Smith et al. designed their challenge points to be "gray" rather than "black and white". According to Pian-Smith et al. they "tried to not create scenarios where the confederate was obviously wrong, so that speaking up would be a 'no brainer'". We created our challenges to be "no brainers" in an attempt to remove the possibility of uncertainty regarding the correct decision. This may have led to an increase in the questioning of our participants with respect to what was "obviously" the wrong thing to do.

Analysis of rationale for behaviour using "think aloud" technique

The 16 instances of delay or failure to challenge provided 19 rationales. Participants were not fore-warned that the senior help might be incompetent

and, given the realistic nature of the scenario, it would seem reasonable to assume that their performance reflected their beliefs and attitudes regarding correct behaviour in the presence of a senior. In addition, if the think-aloud allowed the participants to share these beliefs accurately, as suggested by their responses to the post-scenario questionnaire, then this supports the validity of the findings.

In 15 instances the participant is aware that the senior is making the wrong decision and yet they do not challenge, or need prompting from other healthcare staff. The most common reason for not speaking up was "assumed hierarchy", i.e. the senior is not questioned simply because they are more senior, rather than perceived to be more experienced (the second most common reason for not speaking up). Our results agree with those from Kobayashi et al. (2006) who found that the two most common influences on challenging decisions amongst US residents were "knowledge/experience/understanding" and "teamwork/professionalism/hierarchy". The results also agree with the study by St Pierre et al. (2012). When asked why they did not challenge, 37% had no answer, 35% admitted to there being a discrepancy between what they knew and what they did, 12% explained that the authority gradient prevented them from speaking up, while 8% stated that attendings routinely violated standard operating procedures (SOPs) without being challenged. Therefore, in the St Pierre study, when respondents are able to provide a reason for not challenging, the most common reason is the assumed hierarchy.

The hierarchical nature of Medicine has been well described (Leape, 1994, Rex et al., 2000, Sexton et al., 2000, Thomas et al., 2003) and leads to an "authority gradient", which acts as a barrier to communication (Cosby and Croskerry, 2004).

Our results are also in agreement with work carried out which asked students to write about lapses in professionalism (Lingard et al., 2001, Ginsburg et al., 2003). The students explain their performance by dissociating from the lapse, either by condescending (which we did not observe in our study) or by

referring to “identity mobility”. The latter occurs when a person may take on two or more roles (e.g. “student” versus “doctor in training”) and the person takes on the subordinate role out of self-preservation or deference Ginsburg et al. (2003).

Post-scenario questionnaire

The majority of participants found it easy to perform the think aloud and to remember what they were thinking. A minority found it difficult for varying reasons, e.g. being distracted by poor performance, “not thinking at all” or a preference to watch the performance once in silence. On balance, we felt that the participants would best remember their thoughts if they were asked about these immediately after the scenario. Unfortunately this did mean that the normal practice of debriefing post-scenario was delayed and it is therefore perhaps not surprising that some of the participants were distracted by performance issues.

Limitations

In the simulation setting there is no chance of patient harm. This is one of the strengths of simulation (see Chapter 4) but it may also impact on the participants’ willingness to challenge as they know that there will be no real harm regardless of the outcome. However, none of the participants mentioned this as a reason in their “think aloud” session, e.g. “I knew it was the wrong decision but I didn’t say anything because I knew it didn’t really matter.”

The simulation setting is realistic but not real. The participants’ average score for “realism” of the scenario was 73 (range 0-100, minimum: 55, maximum 95, SD: ±12). It is therefore possible that the participants’ behaviour and actions did not reflect “real-life” performance. Pian-Smith et al. (2009) found that participants were less likely to challenge when the situation was time-critical. Unfortunately the reliable, repeatable creation of such events is probably only possible in the simulator for the foreseeable future.

Ericsson and Simon (1980) distinguished between two types of verbal report: concurrent and retrospective. The think aloud technique we used was a type of retrospective verbal report. The participants were prompted to maintain a flow of monologue, in part to prevent confabulation or mis-remembering. However, it is possible that participants' ability to remember what they were thinking was impaired. We decided against using a concurrent verbal report for three reasons. Firstly, the cognitive load required of the participants to speak aloud what they were thinking while they were dealing with a stressful emergency was considered too onerous. Secondly, the realism of the scenario would have been degraded by the participants' un-natural monologue. Lastly, Fonteyn et al. (1993) state that the retrospective verbal report "might provide inconsistent or incomplete information about one's thinking during a specific problem-solving task, although it could provide a more complete description about one's reasoning strategies." We were interested in not only what the participants were thinking but also how they made their decisions.

Conclusion

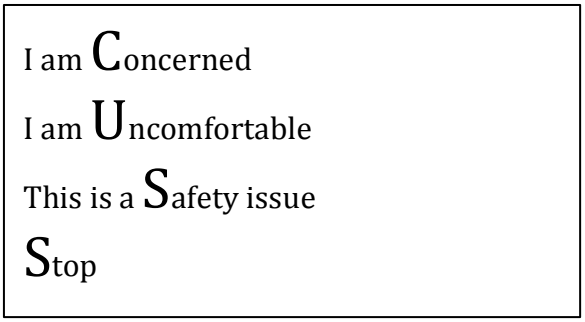
The hazards posed by authority gradients are not unique to healthcare. In his landmark study on obedience to authority, Milgram (1963) persuaded volunteers to "electrocute" an assistant with increasingly powerful shocks. 65% of volunteers continued until the end of the experiment, even though "some believed they had actually killed the other participant" (p.48) (Rees and Knight, 2007). Authority gradients are also found in the maritime industry (Bocanete and Hanzu-Pazara, 2005) and in aviation (Alkov et al., 1992, Gupta, 2004). The worst civilian aviation disaster, excluding the 11th September 2009 terrorist attacks in the USA, was the collision of two Boeing 747s on the runway at Tenerife airport in 1977. A failure by the First Officer to question the actions of the Captain, KLMs most senior training pilot, resulted in the deaths of 583 people (Whittingham, 2004). However, while aviation has moved on to embrace shallow authority gradients, some medical specialties continue to oppose them (Sexton et al., 2000).

The hierarchical structure of medicine has endured since the 19th century (Walton, 2006), with medical students at the bottom and consultants and professors at the top (see Chapter 2: Focus groups: Theme 2: "The hidden curriculum"). Walton (2006) argues that the hierarchy served well in an apprenticeship model of training but that the current system has become a power dynamic between a superior and a subordinate. This unhealthy relationship leads to obsequious students and trainees failing to challenge the more senior doctors. Mahlmeister (2005) meanwhile describes effective team communication as nonhierarchical: "All members of the team have an obligation to speak up; all members of the team have an obligation to listen" (p.296).

Unfortunately, the abolition of authority gradients is not the solution, as some degree of authority gradient is essential for teams to be effective (Australian Bureau of Air Safety Investigation (BASI), 1991). One can differentiate however between the formal authority gradient, as expressed in the difference in seniority between two people, and the informal authority gradient which

depends on behaviour, leadership style and communication (Grech et al., 2008). In aviation, the recognition of the importance of teamwork and team communication has led to the development of crew resource management (CRM) training (Helmreich, 2000). CRM includes techniques for flattening the informal authority gradient as well as techniques for challenging against an authority gradient. The latter include specific phrases to use when challenging a decision or behaviour (Figure 5-1).

Figure 5-1: The CUSS challenge technique



Although unsafe authority gradients are not unique to healthcare, their pervasive nature in healthcare causes extensive morbidity and mortality (Chassin and Becher, 2002, Sutcliffe et al., 2004, Sachs, 2005). Our results suggest that the deference to authority, despite an appreciation that the senior is making a mistake, persists in medical undergraduates. Effective team communication includes the willingness to speak up against authority gradients (Duffy et al., 2004) and embedding CRM principles, such as effective teamworking, in the medical undergraduate curriculum might lead to significant improvement in the willingness of students to challenge poor behaviour. In addition, the implementation of these techniques by undergraduates, in realistic scenarios, could be assessed using a tool such as that presented in Chapter 4.

Chapter 6: Conclusion

Conclusion

The current Western educational system encourages individual excellence (Chakraborti et al., 2008). As students move through from primary to secondary and then university-based education the primacy of individual effort and achievement is emphasised. However, when the undergraduate medical student becomes a doctor he/she becomes part of a complex, dynamic system where effective teamwork is essential. Despite the need for teamwork, even in the postgraduate arena there is a strong tendency to work in uni-professional silos (Kohn et al., 2000, Khalili et al., 2014). The argument for the need to embed teamwork and leadership teaching and evaluation throughout the undergraduate curriculum has been made (O'Sullivan et al., 2012).

The existence of the hidden curriculum has been well-documented (Phillips and Clarke, 2012). O'Sullivan et al. (2012) argue that "the traditional medical school climate of humiliation, competition and hierarchy is an obstacle to learning" (p.e70). Undergraduates, who are at an early stage of their professional development (Hilton and Slotnick, 2005), should not be held to a higher standard than their postgraduate seniors. The lack of a "formal professional continuum" (van Mook et al., 2009b) is a factor in the unprofessional behaviour witnessed by the focus group participants, e.g. calling a student a Nazi because of her German name and naming a student after a colostomy bag. As Irvine (1997), former president of the GMC, stated: "(t)he everyday behaviour of clinical teachers is the living demonstration of their expertise, ethics, and commitment: their professionalism" (p.1542). This everyday behaviour must be made to align with the standards expected of role models, whose attitudes and actions have a disproportionate influence on undergraduates (Byszewski et al., 2012, Morihara et al., 2013, Wong and Trollope-Kumar, 2014). As Glavin and Maran (2003) state: "All of these efforts will be to little avail if they are not reinforced either directly or indirectly via role models in the real clinical setting" (p.63). In addition, it is also the hidden curriculum which will teach undergraduates professional behaviours such as communication and interpersonal skills if the formal curriculum does not accept the challenge (Duffy et al., 2004, van Mook et al., 2009d). The hidden curriculum need not be

entirely negative, as Phillips and Clarke (2012) argue, when teaching "is particularly inspiring, students notice and may be influenced to the extent that they rethink personal beliefs and plans to fit their future doctor selves to these models" (p.887).

As concluded in Chapter 4, the assessment tool is feasible, acceptable and can be cost-effective. Agreement on a final list of elements would benefit from a Delphi process and additional psychometric data are still required, however the shift in assessment theory and practice away from the primacy of psychometrics (Hodges, 2013) may mean that these data will be considered less important than, for example, considerations of feasibility and cost-effectiveness.

However, it is likely that summative, rather than formative, assessment is a more powerful driver of learning (Raupach et al., 2013). With the appreciation that multiple sampling is needed, using this type of evaluation tool for both summative and formative assessment may reify a number of benefits, including a matching up between the goals for learning and the content of the assessment (Duffy et al., 2004).

It is accepted that assessment drives learning, and rather than complaining that students will only learn what we assess, we should make the assessments relevant to students (Schuwirth and van der Vleuten, 2010). This will be more easily achieved if we ensure that the assessments are linked to real-world, applied performance. Schuwirth and Ash (2013) support this claim by arguing for a holistic assessment of competence. In addition, our focus group work supported the need for giving students responsibilities for patient care. This concept has been endorsed by others (Eley and Stallman, 2014). The use of immersive simulation provides students with the context in which they can exercise their skills and receive feedback on their strengths and weaknesses (Schuwirth and van der Vleuten, 2011). This feedback on realistic behaviour may, in turn, lead to the catalytic effect of positive behavioural and attitudinal changes discussed by Norcini et al. (2011).

Wilkinson et al. (2009) identified nine categories of professionalism assessment tools. The evidence firmly supports multi-modal, quantitative and qualitative, multi-agency assessment of the individual within the team (Schuwirth and Van Der Vleuten, 2004, van der Vleuten and Schuwirth, 2005, Goldie, 2013). This tool may shed light on performance which is not easily observed, or assessed, elsewhere in the undergraduate's training. Additional recommendations are provided below.

Recommendations

1. Undergraduates should be provided with the knowledge-base regarding non-technical skills, including teamwork, leadership, authority gradients and human performance limitations, at an early stage (van Mook et al., 2009a, Ginsburg and Lingard, 2011, Hodges et al., 2011). As Glavin and Maran (2003) point out, leadership skills such as resource utilisation and task delegation, are relevant to both clinical practice and to the efficient groupwork required with problem-based learning.
2. The possibility of using immersive simulation-based assessment as final assessments of clinical competence should be explored. Although OSCEs have been called the "gold standard for clinical assessment" (Norman, 2002), they have also been criticised for measuring clinical skills in isolation (van der Vleuten, 2000), for their artificiality (Arnold, 2002) and for lack of correlation with residency-director evaluation scores (Gaur and Skochelak, 2004). According to Eva (Eva and Hodges, 2012), Harden meant for the OSCE "to ensure that students would be observed performing clinical tasks". However, given the context-specificity of behaviour, simulation-based assessment is more likely to allow students to demonstrate "shows how" levels of competence (St Pierre et al., 2012). Many students still feel unprepared for their first posts (Evans and Roberts, 2006) and it is possible that a lack of appropriate evaluation/assessment contributes to this feeling.
3. Further research is required to evaluate the different types of leadership and teamwork expected of undergraduates. The current tool was used in a "crisis" setting. As Shumway (2004) states: "Different leadership is needed for different situations" (p.398) and leadership requirements of the manager/doctor may be very different (Till et al., 2014).
4. Further research is required into the relationship between non-technical and technical skills. DiMatteo and DiNicola (1981) found a strong correlation between the two, while Haurani et al. (2007) showed an association between low interpersonal skills and communication scores and low medical knowledge scores. Further evidence may dispel the

myth of the unprofessional but technically gifted physician referred to in the focus groups in Chapter 3.

Endnote

The Preface referred to two events with very different outcomes. Captain Sullenberger and his crew, through effective leadership and teamwork, saved the lives of their passengers. Elaine Bromiley's medical team, through a lack of leadership and teamwork, were, in part, responsible for her death. The implication is not that aviation is replete with heroes and healthcare workers are villains. Both industries employ fallible human beings who are, at times, fatigued, angry, clumsy or forgetful. However two major differences exist between aviation and healthcare. First, the aviation industry has adopted a "safety culture". The safety culture:

- encourages reporting of incidents,
- is "just" in its response to violations and accidents and
- aims to learn from mistakes.

Secondly, the aviation industry has embraced the teaching, practicing and high-stakes assessment of non-technical skills such as leadership and teamwork. Aircrew are evaluated twice-yearly on their technical and non-technical skills and a failure in either can lead to loss of flying privileges. However, the assessments also serve to provide detailed feedback and to identify retraining, unlike much of the high-stakes assessment prevalent in Medicine (Flin et al., 2003).

Healthcare workers are not pilots and patients are not aeroplanes. However until medical schools, deaneries, health boards, medical indemnity organisations and the GMC encourage a safety culture and insist on the training in, and "just" assessment of, non-technical skills, failures in leadership and teamwork will continue to result in preventable morbidity and mortality. This assessment tool may allow us to evaluate and promote the behaviours, including teamwork and leadership, found to be lacking by the many inquiries.

Research outputs

O'SULLIVAN, H., MONEYPENNY, M. & MCKIMM, J. 2015. Leading and Working in Teams. *British Journal of Hospital Medicine,* 76, 264-269.

MONEYPENNY, M., GUHA, A., MERCER, S., O'SULLIVAN, H. & MCKIMM, J. 2013. Don't follow your leader: challenging erroneous decisions. *British Journal of Hospital Medicine,* 74**,** 687-690.

O'SULLIVAN, H., GUHA, A. & MONEYPENNY, M. 2013. Assessing leadership skills in medical undergraduates. *In:* KER, J. Ch13 Simulation in practice, In: FORREST, K., MCKIMM, J. & EDGAR, S. (eds.) *Essential Simulation in Clinical Education.* Oxford, UK: Wiley-Blackwell.

A personal learning journey

In 2008 I watched a short video called "Just a routine operation" (Clinical Human Factors Group, 2008) and experienced something of a revelation. The video, referred to in the Preface, was a portrayal of the death of Elaine Bromiley. A team of experienced healthcare professionals failed to recover from an initial problem of securing Elaine's airway after she had been anaesthetised. The video showed a chain of errors, failures in teamwork and inter-professional working and uncompensated individual weaknesses.

The revelation for me was that despite individual excellence, which is what medical schools and post-graduate exams select for, patients still die when the team fails to function. As Lingard (2009) stated: "competent individual professionals can—and do, with some regularity— combine to create an incompetent team".

My research background had been positivist and quantitative. My degree in Biochemistry involved laboratory-based work on voltage-gated potassium channels and reverse transcription of cocoa bean enzymes. At medical school, audits and small research projects often involved looking at data from blood tests or X-rays. As a trainee anaesthetist, quality improvement projects and audits included additional quantitative studies such as post-operative nausea, and endotracheal cuff pressures.

The Bromiley video had opened my eyes to the so-called "soft" skills, such as communication, teamwork and leadership. I therefore jumped at the opportunity to take up a fellowship in medical education at the Centre for Excellence in Developing Professionalism (CEDP), with a focus on teamwork and leadership in medical undergraduates. The qualitative aspects of the research, such as the focus groups, meant a steep learning curve.

The metaphorical climb was made easier by discussions with my supervisors and a number of other researchers at the CEDP, including Ray Fewtrell, Peter Leadbetter, Jayne Garner and Simon Watmough. I also found a number of books useful, including:

- Doing Focus Groups (Barbour, 2007),
- Doing qualitative research (Crabtree and Miller, 1992),
- Focus groups as qualitative research (Morgan, 1997b),
- Qualitative research and evaluation methods (Patton, 2002) and
- Understanding medical education: Evidence, theory and practice (Swanwick, 2010)

This new appreciation for the benefits and pitfalls of qualitative research forced me to critically appraise not only the focus group data, but also the data from the simulation-based assessment part of the study. Appreciating that the focus groups merely allow us to see a facet of the undergraduate medical experience, while attempting to quantify social constructs such as leadership and teamwork with an assessment tool, created a degree of tension within the project (and the researcher).

More recently, as director of the Scottish Centre for Simulation and Clinical Human Factors, I am involved in over-seeing research projects. The nature of simulation training, with a focus on improving performance through changing behaviours, means that the understanding I gained during my fellowship years continues to be applicable to day-to-day activities.

References

ACADEMY OF MEDICAL ROYAL COLLEGES 2009. Improving Assessment. London.

ALBRECHT, T. L., JOHNSON, G. M. & WALTHER, J. B. 1993. Understanding Communication Processes in Focus Groups. *In:* MORGAN, D. L. (ed.) *Successful focus groups: advancing the state of the art.* London: SAGE Publications.

ALIMO-METCALFE, B. & ALBAN-METCALFE, J. 2006. More (good) leaders for the public sector. *International Journal of Public Sector Management,* 19**,** 293-315.

ALKOV, R. A., BOROWSKY, M. S., WILLIAMSON, D. W. & YACAVONE, D. W. 1992. The effect of trans-cockpit authority gradient on Navy/Marine helicopter mishaps. *Aviation, space, and environmental medicine*.

AMERICAN BOARD OF INTERNAL MEDICINE 2002. Medical professionalism in the new millennium: a physician charter. *Annals of Internal Medicine,* 136**,** 243.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION 1999. *Standards for educational and psychological testing,* Washington, DC, American Educational Research Association.

ARNOLD, L. 2002. Assessing professional behavior: yesterday, today, and tomorrow. *Academic Medicine,* 77**,** 502-515.

ARNOLD, L. & STERN, D. T. 2006. What is medical professionalism? *In:* STERN, D. T. (ed.) *Measuring Medical Professionalism.* Oxford: Oxford University Press.

ASBURY, J.-E. 1995. Overview of focus group research. *Qualitative health research,* 5**,** 414-420.

ASSOCIATED PRESS. 2009. *Hudson River pilot: 'We were just doing our jobs'* [Online]. The Independent. Available: http://www.independent.co.uk/news/world/americas/hudson-river-pilot-we-were-just-doing-our-jobs-1515488.html [Accessed 28th October 2014].

AUSTRALIAN BUREAU OF AIR SAFETY INVESTIGATION (BASI) 1991. Puma SA 330J Helicopter VH-WOF, Mermaid Sound, Western Australia, 12 May 1991.

BAHAZIQ, W. & CROSBY, E. 2011. Physician professional behaviour affects outcomes: a framework for teaching professionalism during anesthesia residency. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie,* 58**,** 1039-1050.

BAKER, D., GUSTAFSON, S., BEAUBIEN, J., SALAS, E. & BARACH, P. 2005a. *Medical teamwork and patient safety: the evidence-based relation*, Citeseer.

BAKER, D., MULQUEEN, C. & DISMUKES, R. 2001. Training raters to assess resource management skills. *In:* SALAS, E., BOWERS, C. A. & EDENS, E. (eds.) *Improving teamwork in organisations: applications of resource management training.* Mahwah, New Jersey, USA: Laurence Erlbaum Associates.

BAKER, D. P. & DISMUKES, R. K. 2002. A framework for understanding crew performance assessment issues. *The International Journal of Aviation Psychology,* 12**,** 205-222.

BAKER, D. P., SALAS, E., KING, H., BATTLES, J. & BARACH, P. 2005b. The role of teamwork in the professional education of physicians: current status and assessment recommendations. *Joint Commission Journal on Quality and Patient Safety,* 31**,** 185-202.

BARBOUR, R. S. 2005. Making sense of focus groups. *Medical Education,* 39**,** 742-750.

BARBOUR, R. S. 2007. Doing Focus Groups. *The SAGE Qualitative Research Kit*.

BARROW, M. 2012. Conflict in context: designing authentic teamwork education. *Medical Education,* 46**,** 926-927.

BASCH, C. E. 1987. Focus group interview: An underutilized research technique for improving theory and practice in health education. *Health Education & Behavior,* 14**,** 411-448.

BASEHORE, P. M., POMERANTZ, S. C. & GENTILE, M. 2014. Reliability and benefits of medical student peers in rating complex clinical skills. *Medical Teacher,* 36**,** 409-414.

BENSFIELD, L., SOLARI-TWADELL, P. A. & SOMMER, S. 2008. The use of peer leadership to teach fundamental nursing skills. *Nurse Educ,* 33**,** 155-8.

BEYEA, S. C. & NICOLL, L. H. 2000a. Collecting, analyzing, and interpreting focus group data. *AORN Journal,* 71**,** 1278-1283.

BEYEA, S. C. & NICOLL, L. H. 2000b. Learn more using focus groups. *AORN journal,* 71**,** 897-900.

BEYEA, S. C. & NICOLL, L. H. 2000c. Methods to conduct focus groups and the moderator's role. *AORN journal,* 71**,** 1067-1068.

BIRDEN, H., GLASS, N., WILSON, I., HARISON, M., USHERWOOD, T. & NASS, D. 2014. Defining professionalism in medical education: A systematic review. *Medical Teacher,* 36**,** 47-61.

BLEAKLEY, A. 2013. Gender matters in medical education. *Medical Education,* 47**,** 59-70.

BLOOR, M. 1997. Techniques of validation in qualitative research: A critical commentary. *In:* MILLER, G. & DINGWALL, R. (eds.) *Context and method in qualitative research.* London: Sage.

BLOOR, M., FRANKLAND, J., THOMAS, M. & ROBSON, K. 2001. *Focus Groups in Social Research,* London, SAGE.

BOCANETE, P. & HANZU-PAZARA, R. The Influence of Team Errors in Maritime Safety. International Conference on Marine Research and Transportation, 2005 Naples, Italy.

BOLTON V THE LAW SOCIETY 1993. The Incorporated Council of Law Reporting. *England and Wales Court of Appeal (Civil Division)*.

BOULET, J., MURRAY, D., KRAS, J., WOODHOUSE, J., MCALLISTER, J. & ZIV, A. 2003. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology,* 99**,** 1270-1280.

BOURDIEU, P. 1996. Understanding. *Theory, Culture & Society,* 13**,** 17-37.

BOURSICOT, K., ETHERIDGE, L., SETNA, Z., STURROCK, A., KER, J., SMEE, S. & SAMBANDAM, E. 2011. Performance in assessment: Consensus

statement and recommendations from the Ottawa conference. *Medical Teacher,* 33**,** 370-383.

BRAINARD, A. H. & BRISLEN, H. C. 2007. Viewpoint: learning professionalism: a view from the trenches. *Academic Medicine,* 82**,** 1010-1014.

BRETT-FLEEGLER, M. B., VINCI, R. J., WEINER, D. L., HARRIS, S. K., SHIH, M.-C. & KLEINMAN, M. E. 2008. A simulator-based tool that assesses pediatric resident resuscitation competency. *Pediatrics,* 121**,** e597-e603.

BRINKMAN, W. B., GERAGHTY, S. R., LANPHEAR, B. P., KHOURY, J. C., DEL REY, J. A. G., DEWITT, T. G. & BRITTO, M. T. 2006. Evaluation of resident communication skills and professionalism: A matter of perspective? *Pediatrics,* 118**,** 1371-1379.

BROADFOOT, P. M. 1996. Assessment for learning: power or partnership? *In:* GOLDSTEIN, H. & LEWIS, T. (eds.) *Assessment: problems, developments and statistical issues.*: John Wiley and Sons, Inc.

BROWN, N. & DOSHI, M. 2006. Assessing professional and clinical competence: the way forward. *Advances in Psychiatric Treatment,* 12**,** 81-89.

BUCKNALL, V. & PYNSENT, P. B. 2009. Sex and the orthopaedic surgeon: a survey of patient, medical student and male orthopaedic surgeon attitudes towards female orthopaedic surgeons. *The Surgeon,* 7**,** 89-95.

BUELOW, J. R., RATHSACK, C., DOWNS, D., JORGENSEN, K., KARGES, J. R. & NELSON, D. 2008. Building interdisciplinary teamwork among allied health students through live clinical case simulations. *Journal of allied health,* 37**,** e109-23.

BUNNISS, S. & KELLY, D. R. 2010. Research paradigms in medical education research. *Medical Education,* 44**,** 358-366.

BUSH, M. C., JANKOUSKAS, T. S., SINZ, E. H., RUDY, S., HENRY, J. & MURRAY, W. B. 2007. A method for designing symmetrical simulation scenarios for evaluation of behavioral skills. *Simulation in Healthcare,* 2**,** 102-109.

BYSZEWSKI, A., HENDELMAN, W., MCGUINTY, C. & MOINEAU, G. 2012. Wanted: role models-medical students' perceptions of professionalism. *BMC medical education,* 12**,** 115.

CAREY, J. W., MORGAN, M. & OXTOBY, M. J. 1996. Intercoder Agreement in Analysis of Responses to Open-Ended Interview Questions: Examples from Tuberculosis Research. *Cultural Anthropology Methods,* 8**,** 1-5.

CAREY, M. A. 1994. The group effect in focus groups: Planning, implementing, and interpreting focus group research. *In:* MORSE, J. (ed.) *Critical issues in qualitative methodology research.* Thousand Oaks, California, USA: Sage.

CAREY, M. A. 1995a. Comment: concerns in the analysis of focus group data. *Qualitative health research,* 5**,** 487-495.

CAREY, M. A. 1995b. Introduction. *Qualitative health research,* 5**,** 413.

CARLINE, J. D. 2004. Funding medical education research: opportunities and issues. *Academic Medicine,* 79**,** 918-924.

CARLSON, J., MIN, E. & BRIDGES, D. 2009. The impact of leadership and team behavior on standard of care delivered during human patient simulation: A pilot study for undergraduate medical students. *Teaching and learning in medicine,* 21**,** 24-32.

CARRACCIO, C., WOLFSTHAL, S. D., ENGLANDER, R., FERENTZ, K. & MARTIN, C. 2002. Shifting paradigms: from Flexner to competencies. *Academic Medicine,* 77**,** 361-367.

CARTHEY, J. 2003. The role of structured observational research in health care. *Quality and safety in health care,* 12**,** ii13-ii16.

CENTRE FOR COGNITIVE AGEING AND COGNITIVE EPIDEMIOLOGY. 2013. *Systematic reviews and meta-analyses: a step-by-step guide* [Online]. Available: http://www.ccace.ed.ac.uk/research/software-resources/systematic-reviews-and-meta-analyses/step4 [Accessed 7th April 2015 2015].

CHAKRABORTI, C., BOONYASAI, R. T., WRIGHT, S. M. & KERN, D. E. 2008. A systematic review of teamwork training interventions in medical student and resident education. *Journal of General Internal Medicine,* 23**,** 846-853.

CHAMBERS, K., BOULET, J. R. & GARY, N. 2000. The management of patient encounter time in a high-stakes assessment using standardized patients. *Medical Education,* 34**,** 813-817.

CHANDRATILAKE, M. 2014. From the professionalism of a profession to the professionalism of a multiprofessional team. *Medical Education,* 48**,** 345-347.

CHARD, D., ELSHARKAWY, A. & NEWBERY, N. 2006. Medical professionalism: the trainees' views. *Clinical medicine,* 6**,** 68-71.

CHASSIN, M. R. & BECHER, E. C. 2002. The wrong patient. *Annals of Internal Medicine,* 136**,** 826-833.

CHRISTENSON, J., PARRISH, K., BARABE, S., NOSEWORTHY, R., WILLIAMS, T., GEDDES, R. & CHALMERS, A. 1998. A comparison of multimedia and standard advanced cardiac life support learning. *Academic emergency medicine,* 5**,** 702-708.

CLARIDGE, J., CALLAND, J., CHANDRASEKHARA, V., YOUNG, J., SANFEY, H. & SCHIRMER, B. 2003. Comparing resident measurements to attending surgeon self-perceptions of surgical educators. *American Journal of Surgery,* 185**,** 323-327.

CLAUSER, B. E., MARGOLIS, M. J. & SWANSON, D. B. 2002. An Examination of the Contribution of Computer‐based Case Simulations to the USMLE Step 3 Examination. *Academic Medicine,* 77**,** S80-S82.

CLINICAL HUMAN FACTORS GROUP. 2008. *Just a Routine Operation teaching video* [Online]. Available: http://chfg.org/articles-films-guides/films/just-a-routine-operation-teaching-video [Accessed 30th October 2014].

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20**,** 37-46.

COHEN, J. J. 2006. Professionalism in medical education, an American perspective: from evidence to accountability. *Medical Education,* 40**,** 607-617.

COHEN, R., ROTHMAN, A. I., POLDRE, P. & ROSS, J. 1991. Validity and generalizability of global ratings in an objective structured clinical examination. *Academic Medicine,* 66**,** 545-548.

COOK, D. A. & BECKMAN, T. J. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine,* 119**,** 166. e7-166. e16.

COOPER, G. S. & MCCLURE, J. 2005. Why Mothers Die 2000-2002: Executive Summary and Key Findings. London: Royal College of Obstetricians and Gynaecologists.

COOPER, S., CANT, R., PORTER, J., SELLICK, K., SOMERS, G., KINSMAN, L. & NESTEL, D. 2010. Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). *Resuscitation,* 81**,** 446-452.

COOPER, S. & WAKELAM, A. 1999. Leadership of resuscitation teams: Lighthouse Leadership. *Resuscitation,* 42**,** 27-45.

COOPER, T. 2002. Professionalism: Doing the right thing. *Annals of Behavioral Sciences and Medical Education,* 8**,** 119-20.

COSBY, K. S. & CROSKERRY, P. 2004. Profiles in patient safety: authority gradients in medical error. *Academic emergency medicine,* 11**,** 1341-1345.

CÔTÉ-ARSENAULT, D. & MORRISON-BEEDY, D. 1999. Practical advice for planning and conducting focus groups. *Nursing Research,* 48**,** 280-283.

COWAN, M. L. & CLOUTIER, M. G. 1988. Medical simulation for disaster casualty management training. *Journal of Trauma-Injury, Infection, and Critical Care,* 28**,** S178-S182.

COX, K. 1990. No Oscar for OSCA. *Medical Education,* 24**,** 540-545.

CRABTREE, B. F. & MILLER, W. L. 1992. *Doing qualitative research,* London, Sage Publications.

CROASMUN, J. T. & OSTROM, L. 2011. Using Likert-Type Scales in the Social Sciences. *Journal of Adult Education,* 40**,** 19-22.

CRONBACH, L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika,* 16**,** 297-334.

CRONBACH, L., NAGESWARI, R. & GLESER, G. 1963. Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology,* 16**,** 137-163.

CROSSLEY, J., HUMPHRIS, G. & JOLLY, B. 2002. Assessing health professionals. *Medical Education,* 36**,** 800-804.

CROTTY, M. 1998. The Foundations of Social Research: Meaning and perspective in the social research process. London: SAGE Publications Ltd.

CRUESS, R. & CRUESS, S. 1997. Professionalism must be taught. *British Medical Journal,* 315**,** 1674-1677.

CRUESS, R., MCILROY, J. H., CRUESS, S., GINSBURG, S. & STEINERT, Y. 2006. The professionalism mini-evaluation exercise: a preliminary investigation. *Academic Medicine,* 81**,** S74-S78.

CRUESS, S. R., CRUESS, R. L. & STEINERT, Y. 2010. Linking the teaching of professionalism to the social contract: a call for cultural humility. *Medical Teacher,* 32**,** 357-359.

CUESTA-BRIAND, B., AURET, K., JOHNSON, P. & PLAYFORD, D. 2014. 'A world of difference': a qualitative study of medical students' views on professionalism and the'good doctor'. *BMC medical education,* 14**,** 77.

CUMMINGS, G. G., MACGREGOR, T., DAVEY, M., WONG, C. A., LO, E., MUISE, M. & STAFFORD, E. 2010. Leadership styles and outcome patterns for the nursing workforce and work environment: a systematic review. *International journal of nursing studies,* 47**,** 363-385.

CUNNINGTON, J. P. W., NEVILLE, A. J. & NORMAN, G. R. 1996. The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education,* 1**,** 227-233.

CUSHING, A. M., KER, J. S., KINNERSLEY, P., MCKEOWN, P., SILVERMAN, J., PATTERSON, J. & WESTWOOD, O. M. R. 2014. Patient safety and communication: A new assessment for doctors trained in countries where language differs from that of the host country: Results of a pilot using a domain-based assessment. *Patient education and counseling,* 95**,** 332-339.

DANNEFER, E. F. 2013. Beyond assessment of learning toward assessment for learning: Educating tomorrow's physicians. *Medical Teacher,* 35**,** 560-563.

DANNEFER, E. F., HENSON, L. C., BIERER, S. B., GRADY-WELIKY, T. A., MELDRUM, S., NOFZIGER, A. C., BARCLAY, C. & EPSTEIN, R. M. 2005. Peer assessment of professional competence. *Medical Education,* 39**,** 713-722.

DARZI, A. 2008. High Quality Care for All. London: Department of Health of the United Kingdom, HMSO.

DAVIS, D. A., MAZMANIAN, P. E., FORDIS, M., VAN HARRISON, R., THORPE, K. E. & PERRIER, L. 2006. Accuracy of Physician Self-assessment Compared With Observed Measures of CompetenceA Systematic Review. *Journal of the American Medical Association,* 296**,** 1094-1102.

DAY, S., GROSSO, L., NORCINI, J. J., BLANK, L., SWANSON, D. B. & HORNE, M. 1990. Residents' perceptions of evaluation procedures used by their training programme. *Journal of general internal medicine,* 5**,** 421-426.

DEPARTMENT OF HEALTH 2001. Learning from Bristol: The Department of Health's Response to the Report of the Public Inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995. London: The Stationery Office.

DEVELLIS, R. 1991. *Scale Development: Theory and Applications,* Newbury Park, CA, USA, Sage Publications, Inc.

DEVITT, J., KURREK, M., COHEN, M., FISH, K., FISH, P., NOEL, A. & SZALAI, J.-P. 1998. Testing Internal Consistency and Construct Validity During Evaluation of Performance in a Patient Simulator. *Anesthesia & Analgesia,* 86**,** 1160-1164.

DEVITT, J. H., KURREK, M. M., COHEN, M. M. & CLEAVE-HOGG, D. 2001. The validity of performance assessments using simulation. *Anesthesiology,* 95**,** 36-42.

DIMATTEO, M. R. & DINICOLA, D. D. 1981. Sources of assessment of physician performance: a study of comparative reliability and patterns of intercorrelation. *Medical care***,** 829-842.

DOLAN, P., COOKSON, R. & FERGUSON, B. 1999. Effect of discussion and deliberation on the public's views of priority setting in health care: focus group study. *British Medical Journal,* 318**,** 916-919.

DOWNING, S. M. 2003. Validity: on the meaningful interpretation of assessment data. *Medical Education,* 37**,** 830-837.

DOWNING, S. M. 2004. Reliability: on the reproducibility of assessment data. *Medical Education,* 38**,** 1006-1012.

DRISKELL, J. & JOHNSTON, J. 1998. Stress exposure training. *In:* CANNON-BOWERS, J. & SALAS, E. (eds.) *Making decisions under stress - Implications for individual and team training.* Washington, DC: American Psychological Association.

DUFFIELD, K. & SPENCER, J. 2002. A survey of medical students' views about the purposes and fairness of assessment. *Medical Education,* 36**,** 879-886.
DUFFY, F. D., GORDON, G. H., WHELAN, G., COLE-KELLY, K. & FRANKEL, R. 2004. Assessing competence in communication and interpersonal skills: the Kalamazoo II report. *Academic Medicine,* 79**,** 495-507.
DYER, C. 2001. Bristol inquiry: Bristol inquiry condemns hospital's "club culture". *British Medical Journal,* 323**,** 181.
DYER, O. 2004. Doctors suspended for removing wrong kidney. *British Medical Journal,* 328**,** 246.
ECHO. 2010. *Patients support GP stuck in work limbo* [Online]. Available: http://www.echo-news.co.uk/news/8284591.Patients_support_GP_stuck_in_work_limbo/ [Accessed 11th July 2014.
EDWARDS, R. K., KELLNER, K. R., SISTROM, C. L. & MAGYARI, E. J. 2003. Medical student self-assessment of performance on an obstetrics and gynecology clerkship. *American journal of obstetrics and gynecology,* 188**,** 1078-1082.
ELEY, D. S. & STALLMAN, H. 2014. Where does medical education stand in nurturing the 3Rs in medical students: Responsibility, resilience and resolve? *Medical Teacher,* 36**,** 835-837.
ELLIS, A. P., HOLLENBECK, J. R., ILGEN, D. R., PORTER, C. O., WEST, B. J. & MOON, H. 2003. Team learning: collectively connecting the dots. *Journal of applied Psychology,* 88**,** 821.
EPPIC. 2009. *About the EPPI-Centre* [Online]. Available: http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=63 [Accessed 30th January 2015 2015].
EPSTEIN, R. M. 2007. Assessment in Medical Education. *New England Journal of Medicine,* 356**,** 387-396.
EPSTEIN, R. M., DANNEFER, E. F., NOFZIGER, A. C., HANSEN, J. T., SCHULTZ, S. H., JOSPE, N., CONNARD, L. W., MELDRUM, S. C. & HENSON, L. C. 2004. Comprehensive assessment of professional competence: the Rochester experiment. *Teaching and learning in medicine,* 16**,** 186-196.
EPSTEIN, R. M. & HUNDERT, E. M. 2002. Defining and assessing professional competence. *Journal of the American Medical Association,* 287**,** 226-235.
ERDE, E. 2008. Professionalism's facets: Ambiguity, ambivalence, and nostalgia. *Journal of Medical Philosophy,* 33**,** 6-26.
ERICSSON, K. A. & SIMON, H. A. 1980. Verbal reports as data. *Psychological review,* 87**,** 215.
EVA, K. W. & HODGES, B. D. 2012. Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education,* 46**,** 914-919.
EVA, K. W. & REGEHR, G. 2011. Exploring the Divergence between Self-Assessment and Self-Monitoring. *Advances in Health Sciences Education, Theory and Practice,* 16**,** 311-329.
EVA, K. W., ROSENFELD, J., REITER, H. I. & NORMAN, G. R. 2004. An admissions OSCE: the multiple mini‐interview. *Medical Education,* 38**,** 314-326.
EVANS, D. E. & ROBERTS, C. M. 2006. Preparation for practice: how can medical schools better prepare PRHOs? *Medical Teacher,* 28**,** 549-552.

EVENING GAZETTE. 2003. *GP in vice shame can keep his job* [Online]. Gazette Live. Available: http://www.gazettelive.co.uk/news/local-news/gp-vice-shame-can-keep-3850324 [Accessed 11th July 2014.
EVIDENCE FOR POLICY AND PRACTICE INFORMATION AND CO-ORDINATING CENTRE 2007. EPPI-Centre Methods for Conducting Systematic Reviews.
FELPS, W., MITCHELL, T. R. & BYINGTON, E. 2006. How, when, and why bad apples spoil the barrel: Negative group members and dysfunctional groups. *Research in organizational behavior,* 27**,** 175-222.
FERN, E. F. 2001. *Advanced focus group research,* Thousand Oaks, CA, USA, Sage publications.
FERNER, R. E. 2008. Medical error: the plane truth. *British Medical Journal,* 337.
FEUDTNER, C., CHRISTAKIS, D. A. & CHRISTAKIS, N. A. 1994. Do clinical clerks suffer ethical erosion? Students' perceptions of their ethical environment and personal development. *Academic Medicine,* 69**,** 670-9.
FLETCHER, G., FLIN, R. & MCGEORGE, P. 2000. Review of Behavioural Marker Systems Anaesthesia. Aberdeen: University of Aberdeen.
FLETCHER, G., FLIN, R. & MCGEORGE, P. 2003a. Preliminary Evaluation of the Prototype Behavioural Marker System for Anaesthetists' Non-Technical Skills (ANTS). Aberdeen: University of Aberdeen.
FLETCHER, G., FLIN, R., MCGEORGE, P., GLAVIN, R., MARAN, N. & PATEY, R. 2003b. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *British Journal of Anaesthesia,* 90**,** 580-588.
FLETCHER, G., FLIN, R., MCGEORGE, P., GLAVIN, R., MARAN, N. & PATEY, R. 2004. Rating non-technical skills: developing a behavioural marker system for use in anaesthesia. *Cogn Tech Work,* 6**,** 165-171.
FLETCHER, G., MCGEORGE, P., FLIN, R. H., GLAVIN, R. J. & MARAN, N. J. 2002. The role of non-technical skills in anaesthesia: a review of current literature. *British Journal of Anaesthesia,* 88**,** 418-429.
FLIN, R. 1996. *Sitting in the hot seat,* Chichester, UK, John Wiley & Sons.
FLIN, R. 2010. Rudeness at work. *British Medical Journal,* 340**,** 2480.
FLIN, R. & MARAN, N. 2004. Identifying and training non-technical skills for teams in acute medicine. *Quality and Safety in Health Care,* 13(Suppl 1)**,** i80-i84.
FLIN, R., MARTIN, L., GOETERS, K.-M., HORMANN, H.-J., AMALBERTI, R., VALOT, C. & NIJHUIS, H. 2003. Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety,* 3**,** 95-117.
FMLM. 2014. *About Us: Faculty of Medical Leadership and Management* [Online]. Available: https://http://www.fmlm.ac.uk/about-us [Accessed 28th October 2014].
FONTEYN, M. E., KUIPERS, B. & GROBE, S. J. 1993. A description of think aloud method and protocol analysis. *Qualitative Health Research,* 3**,** 430-441.
FOWELL, S. L., MAUDSLEY, G., MAGUIRE, P., LEINSTER, S. J. & BLIGH, J. 2000. Student assessment in undergraduate medical education in the United Kingdom, 1998. *Medical Education,* 34**,** 1-49.
FRAIND, D. B., SLAGLE, J. M., TUBBESING, V. A., HUGHES, S. A. & WEINGER, M. B. 2002. Reengineering intravenous drug and fluid administration processes in the operating room: step one: task analysis of existing processes. *Anesthesiology,* 97**,** 139-147.

FRANCIS, R. 2013. *Report of the Mid Staffordshire NHS foundation trust public inquiry,* London, The Stationery Office.
FRANK, J. R. 2005. *The CanMEDS 2005 physician competency framework: Better standards, better physicians, better care*, Royal College of Physicians and Surgeons of Canada.
FRANKEL, A. S., LEONARD, M. W. & DENHAM, C. R. 2006. Fair and just culture, team behavior, and leadership engagement: the tools to achieve high reliability. *Health services research,* 41**,** 1690-1709.
FREDERIKSEN, N. 1984. The real test bias: Influences of testing on teaching and learning. *American Psychologist,* 39**,** 193.
FREY, J. H. & FONTANA, A. 1993. The Group Interview in Social Research. *In:* MORGAN, D. L. (ed.) *Successful focus groups: advancing the state of the art.* London: SAGE Publications.
GABA, D. M. 2004. The future vision of simulation in health care. *Quality and Safety in Health Care,* 13**,** i2-i10.
GABA, D. M., HOWARD, S. K., FISH, K. J., SMITH, B. E. & SOWB, Y. A. 2001. Simulation-based training in anesthesia crisis resource management (ACRM): a decade of experience. *Simulation & Gaming,* 32**,** 175-193.
GABA, D. M., HOWARD, S. K., FLANAGAN, B., SMITH, B. E., FISH, K. J. & BOTNEY, R. 1998. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology,* 89**,** 8-18.
GARLAND, R. 1991. The mid-point on a rating scale: Is it desirable. *Marketing Bulletin,* 2**,** 66-70.
GAUR, L. & SKOCHELAK, S. 2004. Evaluating Competence in Medical Students. *Journal of the American Medical Association,* 291**,** 2143.
GENERAL MEDICAL COUNCIL 1993. Tomorrow's Doctors. London: GMC.
GENERAL MEDICAL COUNCIL 1995. Good Medical Practice. London: GMC.
GENERAL MEDICAL COUNCIL 2003. Tomorrow's Doctors. London: GMC.
GENERAL MEDICAL COUNCIL 2006a. Good Medical Practice. London: GMC.
GENERAL MEDICAL COUNCIL 2006b. Management for Doctors. London: GMC.
GENERAL MEDICAL COUNCIL 2009. Tomorrow's doctors. London: GMC.
GENERAL MEDICAL COUNCIL 2011. Assessment in undergraduate medical education: Advice supplementary to Tomorrow's Doctors. London: GMC.
GENERAL MEDICAL COUNCIL 2012. Leadership and management for all doctors. London: GMC.
GENERAL MEDICAL COUNCIL 2013. Good Medical Practice. London: GMC.
GENERAL MEDICAL COUNCIL. 2014. *Our role in medical education and training* [Online]. Available: http://www.gmc-uk.org/education/our_role_in_medical_education.asp [Accessed 23rd October 2014].
GENERAL MEDICAL COUNCIL AND MEDICAL SCHOOLS COUNCIL. 2009. *Medical students: professional values and fitness to practise* [Online]. Available: http://www.gmc-uk.org/education/undergraduate/professional_behaviour.asp.
GEORGIOU, A. & LOCKEY, D. J. 2010. The performance and assessment of hospital trauma teams. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine,* 18**,** 66-72.

GERRITY, M. S. & MAHAFFY, J. 1998. Evaluating change in medical school curricula: how did we know where we were going? *Academic medicine,* 73**,** S55-S59.

GILFOYLE, E., GOTTESMAN, R. & RAZACK, S. 2007. Development of a leadership skills workshop in paediatric advanced resuscitation. *Medical Teacher,* 29**,** e276-e283.

GILLESPIE, R., FLORIN, D. & GILLAM, S. 2004. How is patient-centred care understood by the clinical, managerial and lay stakeholders responsible for promoting this agenda? *Health Expectations,* 7**,** 142-148.

GINSBURG, S. & LINGARD, L. 2011. 'Is that normal? Pre‐clerkship students' approaches to professional dilemmas. *Medical Education,* 45**,** 362-371.

GINSBURG, S., REGEHR, G., HATALA, R., MCNAUGHTON, N., FROHNA, A., HODGES, B., LINGARD, L. & STERN, D. 2000. Context, conflict, and resolution: a new conceptual framework for evaluating professionalism. *Academic Medicine,* 75**,** S6-S11.

GINSBURG, S., REGEHR, G. & LINGARD, L. 2003. The disavowed curriculum: Understanding students' reasoning in professionally challenging situations. *Journal of general internal medicine,* 18**,** 1015-1022.

GINSBURG, S., REGEHR, G. & LINGARD, L. 2004. Basing the evaluation of professionalism on observable behaviors: a cautionary tale. *Academic Medicine,* 79**,** S1-S4.

GISEV, N., BELL, J. S. & CHEN, T. F. 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social & Administrative Pharmacy,* 9**,** 330-338.

GLASER, B. 1965. The Constant Comparative Method of qualitative analysis. *Social Problems,* 12**,** 436-445.

GLAVIN, R. & MARAN, N. 2003. Integrating human factors into the medical curriculum. *Medical Education,* 37(Suppl. 1)**,** 59-64.

GOLDIE, J. 2013. Assessment of professionalism: A consolidation of current thinking. *Medical Teacher,* 35**,** e952-e956.

GOLDSMITH, T. E. & JOHNSON, P. J. 2002. Assessing and improving evaluation of aircrew performance. *The International Journal of Aviation Psychology,* 12**,** 223-240.

GOVAERTS, M. J. B., VAN DER VLEUTEN, C. P. M. & SCHUWIRTH, L. W. T. 2002. Optimising the reproducibility of a performance-based assessment test in midwifery education. *Advances in Health Sciences Education,* 7**,** 133-145.

GRAHAM, J., HOCKING, G. & GILES, E. 2010. Anaesthesia non-technical skills: can anaesthetists be trained to reliably use this behavioural marker system in 1 day? *British Journal of Anaesthesia,* 104**,** 440-445.

GRECH, M., HORBERRY, T. & KOESTER, T. 2008. *Human factors in the maritime domain,* Boca Raton, Florida, USA, CRC Press.

GRIFFIN, E. 1997. Groupthink of Irving Janus. *A First Look at Communication Theory.* New York: McGraw-Hill.

GROGAN, E. L., STILES, R. A., FRANCE, D. J., SPEROFF, T., MORRIS JR, J. A., NIXON, B., GAFFNEY, F. A., SEDDON, R. & PINSON, C. W. 2004. The impact of aviation-based teamwork training on the attitudes of health-care professionals. *Journal of the American College of Surgeons,* 199**,** 843-848.

GUPTA, A. 2004. Trans-Cockpit Authority Gradient in Flying Training: A Case Report. *Indian Journal of Aerospace Medicine,* 48**,** 41-4.
GUPTA V GENERAL MEDICAL COUNCIL 2002. The Incorporated Council of Law Reporting. *Privy Council,* [2002] 1 WLR 1691, 1702.
HAFFERTY, F. 2002. What medical students know about professionalism. *Mount Sinai Journal of Medicine,* 69**,** 385-397.
HAFFERTY, F. W. 2006. Measuring medical professionalism: A commentary. *In:* STERN, D. T. (ed.) *Measuring medical professionalism.* Oxford: Oxford University Press.
HALL, P. & WEAVER, L. 2001. Interdisciplinary education and teamwork: a long and winding road. *Medical Education,* 35**,** 867-875.
HAMMAN, W. R. 2004. The complexity of team training: what we have learned from aviation and its applications to medicine. *Quality and Safety in Health Care,* 13**,** i72-i79.
HARMER, M. 2007. *Independent Review on the care given to Mrs Elaine Bromiley on 29 March 2005.* [Online]. Clinical Human Factors Group. Available: http://www.chfg.org/resources/07_qrt04/Anonymous_Report_Verdict_and_Corrected_Timeline_Oct_07.pdf [Accessed 30th October 2014].
HAURANI, M. J., RUBINFELD, I., RAO, S., BEAUBIEN, J., MUSIAL, J. L., PARKER, A., REICKERT, C., RAAFAT, A. & SHEPARD, A. 2007. Are the communication and professionalism competencies the new critical values in a resident's global evaluation process? *Journal of surgical education,* 64**,** 351-356.
HAWKINS, R. E., KATSUFRAKIS, P. J., HOLTMAN, M. C. & CLAUSER, B. E. 2009. Assessment of medical professionalism: Who, what, when, where, how, and … why? . *Medical Teacher,* 31**,** 348-361.
HELMREICH, R. L. 1997. Managing human error in aviation. *Scientific American,* 276**,** 62-67.
HELMREICH, R. L. 2000. On error management: lessons from aviation. *British Medical Journal,* 320**,** 781-785.
HERBERT, P., MESLIN, E. & DUNN, E. 1992. Measuring the ethical sensitivity of medical students: a study at the University of Toronto. *Journal of medical ethics,* 18**,** 142-147.
HICKS, L. K., LIN, Y., ROBERTSON, D. W., ROBINSON, D. L. & WOODROW, S. I. 2001. Understanding the clinical dilemmas that shape medical students' ethical development: questionnaire survey and focus group study. *British Medical Journal,* 322**,** 709-710.
HILTON, S. R. & SLOTNICK, H. B. 2005. Proto-professionalism: how professionalisation occurs across the continuum of medical education. *Medical Education,* 39**,** 58-65.
HJORTDAHL, M., RINGEN, A., NAESS, A.-C. & WISBORG, T. 2009. Leadership is the essential non-technical skill in the trauma team - results of a qualitative study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine,* 17**,** 48.
HO, M. J., LIN, C. W., CHIU, Y. T., LINGARD, L. & GINSBURG, S. 2012. A cross‐cultural study of students' approaches to professional dilemmas: sticks or ripples. *Medical Education,* 46**,** 245-256.
HODGES, B. 2013. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher,* 35**,** 564-568.

HODGES, B. D., GINSBURG, S., CRUESS, R., CRUESS, S., DELPORT, R., HAFFERTY, F. W., HO, M.-J., HOLMBOE, E. S., HOLTMAN, M. C., OHBU, S., REES, C., TEN CATE, O., TSUGAWA, Y., VAN MOOK, W., WASS, V., WILKINSON, T. J. & WADE, W. 2011. Assessment of professionalism: Recommendations from the Ottawa 2010 Conference. *Medical Teacher,* 33**,** 354-363.

HOFF, W., REILLY, P., ROTONDO, M., DIGIACOMO, C. & SCHWAB, W. 1997. The Importance of the Command-Physician in Trauma Resuscitation. *Journal of Trauma-Injury Infection & Critical Care,* 43**,** 772-777.

HOFMANN, B. 2009. Why simulation can be efficient: on the preconditions of efficient learning in complex technology based practices. *BMC medical education,* 9**,** 48.

HOLCOMB, J. B., DUMIRE, R. D., CROMMETT, J. W., STAMATERIS, C. E., FAGERT, M. A., CLEVELAND, J. A., DORLAC, G. R., DORLAC, W. C., BONAR, J. P. & HIRA, K. 2002. Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training. *Journal of Trauma-Injury, Infection, and Critical Care,* 52**,** 1078-1086.

HOLLY, C. & SALMOND, S. W. 2011. *Comprehensive systematic review for advanced nursing practice*, Springer Publishing Company.

HOLMBOE, E. 2004a. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Academic Medicine,* 79**,** 16-22.

HOLMBOE, E. S. 2004b. Faculty and the Observation of Trainees' Clinical Skills: Problems and Opportunities. *Academic Medicine,* 79**,** 16-22.

HOLZMAN, R. S., COOPER, J. B., GABA, D. M., PHILIP, J. H., SMALL, S. D. & FEINSTEM, D. 1995. Anesthesia crisis resource management: real-life simulation training in operating room crises. *Journal of clinical anesthesia,* 7**,** 675-687.

HOWARD, N. M., DOTTL, S. L. & PRUCHA, C. E. 1999. Development of a Leadership-skills-Assessment Instrument for Medical Educators. *Academic Medicine,* 74**,** 609-610.

HOWATSON-JONES, I. L. 2007. Dilemmas of focus group recruitment and implementation: a pilot perspective. *Nurse Researcher,* 14**,** 11.

HUGHES, C., TOOHEY, S. & VELAN, G. 2008. eMed Teamwork: a self-moderating system to gather peer feedback for developing and assessing teamwork skills. *Medical Teacher,* 30**,** 5-9.

IPSOS MORI. 2011. *Ipsos MORI Veracity Index* [Online]. Available: http://www.ipsos-mori.com/Assets/Docs/Polls/Veracity2011.pdf [Accessed 8th July 2014.

IRVINE, D. 1997. The performance of doctors. I: Professionalism and self regulation in a changing world. *British Medical Journal,* 314**,** 1540.

ISSENBERG, B. S., MCGAGHIE, W. C., PETRUSA, E. R., LEE GORDON, D. & SCALESE, R. J. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review*. *Medical Teacher,* 27**,** 10-28.

ISSENBERG, B. S. & SCALESE, R. J. 2007. Best evidence on high‐fidelity simulation: what clinical teachers need to know. *The Clinical Teacher,* 4**,** 73-77.

JAMES, L., DEMAREE, R. & WOLF, G. 1984. Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology,* 69**,** 85-98.

JAMIESON, S. 2004. Likert scales: how to (ab) use them. *Medical Education,* 38**,** 1217-1218.
JEN, M. H., BOTTLE, A., MAJEED, A., BELL, D. & AYLIN, P. 2009. Early in-hospital mortality following trainee doctors' first day at work. *PLoS One,* 4**,** e7103.
JHA, V., BEKKER, H., DUFFY, S. & ROBERTS, T. 2006. Perceptions of professionalism in medicine: a qualitative study. *Medical Education,* 40**,** 1027-1036.
JHA, V., BEKKER, H. L., DUFFY, S. R. & ROBERTS, T. E. 2007. A systematic review of studies assessing and facilitating attitudes towards professionalism in medicine. *Medical Education,* 41**,** 822-829.
JONES, R., PANDA, M. & DESBIENS, N. 2008. Internal medical residents do not accurately assess their medical knowledge. *Advances in Health Sciences Education,* 13**,** 463-468.
KAPFERER, J.-N. 2013. *Rumors: Uses, interpretations, and images,* New Brunswick, NJ, USA, Transaction Publishers.
KAYE, W. & MANCINI, M. E. 1986. Use of the Mega Code to evaluate team leader performance during advanced cardiac life support. *Critical care medicine,* 14**,** 99-104.
KECK, J. W., ARNOLD, L., WILLOUGHBY, L. & CALKINS, V. 1979. Efficacy of cognitive/noncognitive measures in predicting resident-physician performance. *Academic Medicine,* 54**,** 759-65.
KEOGH, B. 2013. *Review into the quality of care and treatment provided by 14 hospital trusts in England: overview report,* London, The Stationery Office.
KER, J., MOLE, L. & BRADLEY, P. 2003. Early introduction to interprofessional learning: a simulated ward environment. *Medical Education,* 37**,** 248-255.
KER, J. S., HESKETH, E. A., ANDERSON, F. & JOHNSTON, D. A. 2006. Can a ward simulation exercise achieve the realism that reflects the complexity of everyday practice junior doctors encounter? *Medical Teacher,* 28**,** 330-334.
KEYNAN, A., FRIEDMAN, M. & BENBASSAT, J. 1987. Reliability of global rating scales in the assessment of clinical competence of medical students. *Medical Education,* 21**,** 477-481.
KHALILI, H., HALL, J. & DELUCA, S. 2014. Historical analysis of professionalism in western societies: implications for interprofessional education and collaborative practice. *Journal of interprofessional care,* 28**,** 92-97.
KHAN, K., PATTISON, T. & SHERWOOD, M. 2011. Simulation in medical education. *Medical Teacher,* 33**,** 1-3.
KIM, J., NEILIPOVITZ, D., CARDINAL, P., CHIU, M. & CLINCH, J. 2006. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Critical care medicine,* 34**,** 2167-2174.
KINZEY, R. 1997. Report Writing Without Guilt. *In:* KRUEGER, R. A. (ed.) *Analyzing and reporting focus group results.* Thousand Oaks, CA, USA: Sage publications.
KIRKPATRICK, D. I. 1998. *Evaluating Training Programs: The Four Levels (2nd ed),* San Francisco, Berrett-Koehler.

KITZINGER, J. 1994. The methodology of Focus Groups: the importance of interaction between research participants. *Sociology of Health & Illness,* 16**,** 19.

KITZINGER, J. 1995. Introducing focus groups. *British Medical Journal,* 311**,** 299-302.

KITZINGER, J. & BARBOUR, R. 1999. *Developing focus group research: politics, theory and practice*, Sage.

KLABER, R. E., ROUECHE, A., HODGKINSON, R. & DAWN CASS, H. 2008. A structured approach to planning a work-based leadership development programme for doctors in training. *The International Journal of Clinical Leadership,* 16**,** 121-129.

KLAMPFER, B., FLIN, R., HELMREICH, R., HAUSLER, R., SEXTON, B., FLETCHER, G., FIELD, P., STAENDER, S., LAUCHE, K., DIECKMANN, P. & AMACHER, A. 2001. Enhancing Performance in High Risk Environments: Recommendations for the use of Behavioural Markers. *In:* KLAMPFER, B. & JOCHUM, K. (eds.) *Group Interaction in High Risk Environments (GIHRE).* Berlin: Humboldt Universitat zu Berlin.

KNEEBONE, R., KIDD, J., NESTEL, D., ASVALL, S., PARASKEVA, P. & DARZI, A. 2002. An innovative model for teaching and learning clinical procedures. *Medical education,* 36**,** 628-634.

KOBAYASHI, H., PIAN-SMITH, M., SATO, M., SAWA, R., TAKESHITA, T. & RAEMER, D. 2006. A cross-cultural survey of residents' perceived barriers in questioning/challenging authority. *Quality and Safety in Health Care,* 15**,** 277-283.

KOHN, L. T., CORRIGAN, J. M. & DONALDSON, M. S. 2000. To Err Is Human: Building a Safer Health System. Washington, DC: National Academy Press.

KREVANS, J. R. & BENSON, J. A. 1983. Evaluation of humanistic qualities in the internist. *Annals of Internal Medicine,* 99**,** 720-724.

KRUEGER, R. A. 1995. The future of focus groups. *Qualitative health research,* 5**,** 524-530.

KRUEGER, R. A. 1997. *Analyzing and reporting focus group results,* Thousand Oaks, CA, USA, Sage publications.

KRUEGER, R. A. & CASEY, M. 1994. *Focus Groups,* Thousand Oaks, CA, USA, Sage Publications.

KRUEGER, R. A. & CASEY, M. 2000. *Focus Groups: A practical guide for applied research (3rd edition),* Thousand Oaks, CA, USA, Sage Publications.

KYRKJEBØ, J. M., BRATTEBØ, G. & SMITH-STRØM, H. 2006. Improving patient safety by using interprofessional simulation training in health professional education. *Journal of interprofessional care,* 20**,** 507-516.

LAW COMMISSION 2012. Joint Consultation Paper LCCP 202 / SLCDP 153 / NILC 12. London: Law Commission.

LEACH, L. S., MYRTLE, R. C., WEAVER, F. A. & DASU, S. 2009. Assessing the performance of surgical teams. *Health care management review,* 34**,** 29-41.

LEAPE, L. L. 1994. Error in medicine. *Journal of the American Medical Association,* 272**,** 1851-1857.

LEDERMAN, L. 1983. High apprehensives talk about communication apprehension and its effects on their behaviour. *Communication Quarterly,* 31**,** 233-237.
LEMIRE, J. A. 2002. Preparing nurse leaders: a leadership education model... originally printed in Nursing Leadership Forum, 6(2). *Nursing Leadership Forum,* 7**,** 47-52.
LEO, T. & EAGEN, K. 2008. Professionalism education: The medical student response. *Perspectives in biology and medicine,* 51**,** 508-516.
LEONARD, M., GRAHAM, S. & BONACUM, D. 2004. The human factor: the critical importance of effective teamwork and communication in providing safe care. *Quality and Safety in Health Care,* 13**,** i85-i90.
LERNER, S., MAGRANE, D. & FRIEDMAN, E. 2009. Teaching Teamwork in Medical Education. *Mount Sinai Journal of Medicine,* 76**,** 318-329.
LEWIS, C. & RIEMAN, J. 1993. Task-centered user interface design. *A Practical Introduction.*
LINGARD, L. 2009. What we see and don't see when we look at 'competence': notes on a god term. *Advances in health sciences education,* 14**,** 625-628.
LINGARD, L. 2014. When I say… grounded theory. *Medical Education,* 48**,** 748-749.
LINGARD, L., ESPIN, S., WHYTE, S., REGEHR, G., BAKER, G., REZNICK, R., BOHNEN, J., ORSER, B., DORAN, D. & GROBER, E. 2004. Communication failures in the operating room: an observational classification of recurrent types and effects. *Quality and Safety in Health Care,* 13**,** 330-334.
LINGARD, L., GARWOOD, K. I. M., SZAUTER, K. & STERN, D. 2001. The rhetoric of rationalization: how students grapple with professional dilemmas. *Academic Medicine,* 76**,** S45-S47.
LINGARD, L. & KENNEDY, T. J. 2007. Qualitative research in medical education. *In:* SWANWICK, T. (ed.) *Understanding Medical Education: Evidence, Theory and Practice.* Edinburgh: Association for the Study of Medical Education.
LITTLE, P., EVERITT, H., WILLIAMSON, I., WARNER, G., MOORE, M., GOULD, C., FERRIER, K. & PAYNE, S. 2001. Preferences of patients for patient centred approach to consultation in primary care: observational study. *British Medical Journal,* 322**,** 468.
LORD MACLEAN 2014. The Vale of Leven Hospital Inquiry Report. Edinburgh: Scottish Parliament.
LUCY, J. A. 1997. Linguistic relativity. *Annual Review of Anthropology,* 26**,** 291-312.
LYNCH, D. C., SURDYK, P. M. & EISER, A. R. 2004. Assessing professionalism: a review of the literature. *Medical Teacher,* 26**,** 366-373.
MAHLMEISTER, L. R. 2005. Preventing adverse perinatal outcomes through effective communication: lessons learned. *The Journal of perinatal & neonatal nursing,* 19**,** 295-297.
MALEC, J. F., TORSHER, L. C., DUNN, W. F., WIEGMANN, D. A., ARNOLD, J. J., BROWN, D. A. & PHATAK, V. 2007. The Mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. *Simulation in Healthcare,* 2**,** 4-10.

MARAN, N. & GLAVIN, R. 2003. Low to high fidelity simulation, a continuum of medical education? *Medical Education,* 37**,** 22-28.

MARSCH, S. C., MULLER, C., MARQUARDT, K., CONRAD, G., TSCHAM, F. & HUNZIKER, P. R. 2004. Human factors affect the quality of cardiopulmonary resuscitation in simulated cardiac arrests. *Resuscitation,* 60**,** 51-56.

MARSHALL, C. & ROSSMAN, G. B. 2010. Managing, Analyzing and Interpreting Data. *Designing Qualitative Research.* London: SAGE Publications Ltd.

MARTIN, J. A., REZNICK, R. K., ROTHMAN, A., TAMBLYN, R. M. & REGEHR, G. 1996. Who should rate candidates in an objective structured clinical examination? *Academic Medicine,* 71**,** 170-5.

MAYS, N. & POPE, C. 1995. Rigour and qualitative research. *British Medical Journal,* 311**,** 109-112.

MAZOR, K., CANAVAN, C., FARRELL, M., MARGOLIS, M. & CLAUSER, B. 2008. Collecting Validity Evidence for an Assessment of Professionalism: Findings from Think-Aloud Interviews. *Academic Medicine,* 83**,** S9.

MCCULLOCH, P., MISHRA, A., HANDA, A., DALE, T., HIRST, G. & CATCHPOLE, K. 2009. The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. *Quality and Safety in Health Care,* 18**,** 109-115.

MCKEGNEY, C. 1989. Medical education: a neglectful and abusive family system. *Family medicine,* 21**,** 452-457.

MCMANUS, I. C., IQBAL, S., CHANDRARAJAN, A., FERGUSON, E. & LEAVISS, J. 2005. Unhappiness and dissatisfaction in doctors cannot be predicted by selectors from medical school application forms: A prospective, longitudinal study. *BMC medical education,* 5**,** 38.

MCNAIR, R. 2005. The case for educating health care students in professionalism as the core content of interprofessional education. *Medical Education,* 39**,** 456-464.

MEDICAL ACT 1983.

MELIA, K. 1997. Producing 'plausible stories': interviewing student nurses. *In:* MILLER, G. & DINGWALL, R. (eds.) *Context and method in qualitative research.* London: Sage.

MENNIN, S. P. & KRACKOV, S. K. 1998. Reflections on relevance, resistance, and reform in medical education. *Academic Medicine,* 73**,** S60-4.

MESSICK, S. 1991. Validity of test interpretation and use. *In:* ALKIN, M. (ed.) *Encyclopaedia of Educational Research (6th ed.).* New York: Macmillan.

MICKAN, S. M. & RODGER, S. A. 2005. Effective health care teams: a model of six characteristics developed from shared perceptions. *Journal of interprofessional care,* 19**,** 358-370.

MILES, M. & HUBERMAN, A. 1994. *An expanded sourcebook: Qualitative data analysis. 2nd ed,* Thousand Oaks, CA, USA, Sage.

MILGRAM, S. 1963. Behavioral study of obedience. *Journal of Abnormal Psychology***,** 371-378.

MILLER, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review,* 63**,** 81-97.

MILLER, G. E. 1990. The assessment of clinical skills/competence/performance. *Academic medicine: journal of the Association of American Medical Colleges,* 65**,** S63-67.

MISHRA, A., CATCHPOLE, K., DALE, T. & MCCULLOCH, P. 2008. The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surgical endoscopy,* 22**,** 68-73.

MITCHELL, L. & FLIN, R. 2008. Non‐technical skills of the operating theatre scrub nurse: literature review. *Journal of advanced nursing,* 63**,** 15-24.

MOINEAU, G., POWER, B., PION, A., WOOD, T. & HUMPHREY-MURTO, S. 2011. Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. *Medical Education,* 45**,** 183-191.

MOORTHY, K., MUNZ, Y., ADAMS, S., PANDEY, V. & DARZI, A. 2005. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. *Annals of surgery,* 242**,** 631-639.

MOORTHY, K., MUNZ, Y., FORREST, D., PANDEY, V., UNDRE, S., VINCENT, C. & DARZI, A. 2006. Surgical crisis management skills training and assessment: a stimulation-based approach to enhancing operating room performance. *Annals of surgery,* 244**,** 139-147.

MOREY, J. C., SIMON, R., JAY, G. D., WEARS, R. L., SALISBURY, M., DUKES, K. A. & BERNS, S. D. 2002. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. *Health services research,* 37**,** 1553-1581.

MORGAN, D. L. 1988. *Focus groups as qualitative research,* Newbury Park, California, USA, Sage.

MORGAN, D. L. 1995. Why things (sometimes) go wrong in focus groups. *Qualitative health research,* 5**,** 516-523.

MORGAN, D. L. 1997a. Computerized Analysis. *In:* KRUEGER, R. A. (ed.) *Analyzing and reporting focus group results.* Thousand Oaks, CA, USA: Sage publications.

MORGAN, D. L. 1997b. *Focus groups as qualitative research,* Thousand Oaks, CA, USA, SAGE Pulications, Inc.

MORGAN, D. L. & KRUEGER, R. A. 1993. When to Use Focus Groups and Why. *In:* MORGAN, D. L. (ed.) *Successful focus groups: advancing the state of the art.* London: SAGE Publications.

MORGAN, P. J., CLEAVE-HOGG, D. & GUEST, C. B. 2001a. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. *Academic Medicine,* 76**,** 1053-1055.

MORGAN, P. J., CLEAVE-HOGG, D. M., GUEST, C. B. & HEROLD, J. 2001b. Validity and reliability of undergraduate performance assessments in an anesthesia simulator. *Canadian Journal of Anesthesia,* 48**,** 225-233.

MORGAN, P. J. & CLEAVE‐HOGG, D. 2000. A Canadian simulation experience: faculty and student opinions of a performance evaluation study. *British Journal of Anaesthesia,* 85**,** 779-781.

MORGAN, P. J., PITTINI, R., REGEHR, G., MARRS, C. & HALEY, M. F. 2007. Evaluating teamwork in a simulated obstetric environment. *Anesthesiology,* 106**,** 907-915.

MORIHARA, S. K., JACKSON, D. S. & CHUN, M. B. J. 2013. Making the professionalism curriculum for undergraduate medical education more relevant. *Medical Teacher,* 35**,** 908-914.
MORISON, S. L. & STEWART, M. C. 2005. Developing interprofessional assessment. *Learning in Health & Social Care,* 4**,** 192-202.
MORRISON-BEEDY, D., CÔTÉ-ARSENAULT, D. & FEINSTEIN, N. F. 2001. Maximizing results with focus groups: Moderator and analysis issues. *Applied Nursing Research,* 14**,** 48-53.
MURRAY, B. W. & FOSTER, P. A. 2000. Crisis Resource Management Among Strangers: Principles of Organizing a Multidisciplinary Group for Crisis Resource Management. *Journal of Clinical Anesthesia,* 12**,** 633-638.
MURRAY, D., BOULET, J., ZIV, A., WOODHOUSE, J., KRAS, J. & MCALLISTER, J. 2002. An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. *Medical Education,* 36**,** 833-841.
MYERSON, K. 1998. Can we assess professional behaviour in anaesthetists? *Anaesthesia,* 53**,** 1039-1040.
NATIONAL ADVISORY GROUP ON THE SAFETY OF PATIENTS IN ENGLAND 2013. A promise to learn - a commitment to act: Improving the Safety of Patients in England. London: The Stationery Office.
NEILY, J., MILLS, P. D., YOUNG-XU, Y., CARNEY, B. T., WEST, P., BERGER, D. H., MAZZIA, L. M., PAULL, D. E. & BAGIAN, J. P. 2010. Association between implementation of a medical team training program and surgical mortality. *Journal of the American Medical Association,* 304**,** 1693-1700.
NEUMANN, M., EDELH√ § USER, F., TAUSCHEL, D., FISCHER, M. R., WIRTZ, M., WOOPEN, C., HARAMATI, A. & SCHEFFER, C. 2011. Empathy decline and its reasons: a systematic review of studies with medical students and residents. *Academic Medicine,* 86**,** 996-1009.
NEWBLE, D. I. & JAEGER, K. 1983. The effect of assessments and examinations on the learning of medical students. *Medical Education,* 17**,** 165-171.
NHS EMPLOYERS. 2009. *European Working Time Directive* [Online]. Available: http://www.nhsemployers.org/your-workforce/need-to-know/european-working-time-directive [Accessed 10th December 2014].
NHS INSTITUTE FOR INNOVATION AND IMPROVEMENT AND ACADEMY OF MEDICAL ROYAL COLLEGES 2010. Medical Leadership Competency Framework (3rd Edition). Coventry.
NHS LEADERSHIP ACADEMY 2013. Healthcare Leadership Model. London: NHS Leadership Academy.
NICHOLS, D. 1998. Choosing an Intraclass Correlation Coefficient. *SPSS Keywords.*
NICOLSON, P. & ANDERSON, P. 2003. Quality of life, distress and self-esteem: A focus group study of people with chronic bronchitis. *British journal of health psychology,* 8**,** 251-270.
NOEL, G. L., HERBERS, J. E., CAPLOW, M. P., COOPER, G. S., PANGARO, L. N. & HARVEY, J. 1992. How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine,* 117**,** 757-765.

NOONAN, L. E. & SULSKY, L. M. 2001. Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance,* 14**,** 3-26.

NORCINI, J. J. 2003. Peer assessment of competence. *Medical Education,* 37**,** 539-543.

NORCINI, J. J., ANDERSON, B., BOLLELA, V., BURCH, V., COSTA, M. J., DUVIVIER, R., GALBRAITH, R., HAYS, R., KENT, A., PERROTT, V. & ROBERTS, T. 2011. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher,* 33**,** 206-214.

NORCINI, J. J. & MCKINLEY, D. W. 2007. Assessment methods in medical education. *Teaching and teacher education,* 23**,** 239-250.

NORMAN, G. 2002. Research in medical education: three decades of progress. *British Medical Journal,* 324**,** 1560-1562.

NURSING & MIDWIFERY COUNCIL 2007. Supporting direct care through simulated practice learning in the pre-registration nursing programme. London: Nursing & Midwifery Council.

NUTTER, D. & WHITCOMB, M. 2001. The AAMC project on the clinical education of medical students. Washington, DC, USA: Association of American Medical Colleges.

O'CONNELL, M. T. & PASCOE, J. M. 2004. Undergraduate medical education for the 21st century: leadership and teamwork. *Family medicine,* 36**,** S51-S56.

O'DANIEL, M. & ROSENSTEIN, A. 2008. Professional Communication and Team Collaboration. *In:* HUGHES, R. G. (ed.) *Patient Safety and Quality: An Evidence-based Handbook for Nurses.* Rockville, Maryland, USA: Agenct for Healthcare Research and Quality.

O'FLYNN, S., KELLY, M. A. & BENNETT, D. 2014. Professionalism and identity formation: students' journeys and emotions. *Medical Education,* 48**,** 463-465.

O'SULLIVAN, H. & MCKIMM, J. 2011a. Doctor as professional and doctor as leader: same attributes, attitudes and values? *British Journal of Hospital Medicine,* 72**,** 463-466.

O'SULLIVAN, H. & MCKIMM, J. 2011b. Medical leadership and the medical student. *British Journal of Hospital Medicine,* 72**,** 346-349.

O'SULLIVAN, H. & MCKIMM, J. 2011c. Medical leadership: an international perspective. *British Journal of Hospital Medicine,* 72**,** 638-641.

O'SULLIVAN, H., VAN MOOK, W., FEWTRELL, R. & WASS, V. 2012. Integrating professionalism into the curriculum: AMEE Guide No. 61. *Medical Teacher,* 34**,** e64-e77.

ODOM, S. L., BRANTLINGER, E., GERSTEN, R., HORNER, R. H., THOMPSON, B. & HARRIS, K. R. 2005. Research in special education: Scientific methods and evidence-based practices. *Exceptional children,* 71**,** 137-148.

OKUDA, Y., BRYSON, E. O., DEMARIA, S., JACOBSON, L., QUINONES, J., SHEN, B. & LEVINE, A. 2009. The Utility of Simulation in Medical Education: What Is the Evidence. *Mount Sinai Journal of Medicine,* 76**,** 330-343.

ORLANDER, J. D., WIPF, J. E. & LEW, R. A. 2006. Development of a tool to assess the team leadership skills of medical residents. *Medical Education Online,* 11**,** 1-6.

ØSTERGAARD, H. T., ØSTERGAARD, D. & LIPPERT, A. 2004. Implementation of team training in medical education in Denmark. *Quality and Safety in Health Care,* 13**,** i91-i95.

OTTESTAD, E., BOULET, J. R. & LIGHTHALL, G. K. 2007. Evaluating the management of septic shock using patient simulation. *Critical care medicine,* 35**,** 769-775.

PAISLEY, A., BALDWIN, P. & PATERSON-BROWN, S. 2005. Accuracy of medical staff assessment of trainees' operative performance. *Medical Teacher,* 27**,** 634-638.

PASKINS, Z. & PEILE, E. 2010. Final year medical students' views on simulation-based teaching: a comparison with the Best Evidence Medical Education Systematic Review. *Medical Teacher,* 32**,** 569-577.

PATEL, P., ROBINSON, B. S., NOVICOFF, W. M., DUNNINGTON, G. L., BRENNER, M. J. & SALEH, K. J. 2011. The disruptive orthopaedic surgeon: implications for patient safety and malpractice liability. *The Journal of Bone & Joint Surgery,* 93**,** e126 1-6.

PATENAUDE, J., NIYONSENGA, T. & FAFARD, D. 2003. Changes in students' moral development during medical school: a cohort study. *Canadian Medical Association Journal,* 168**,** 840-844.

PATTON, M. Q. 2002. *Qualitative research and evaluation methods (3rd ed.),* Thousand Oaks, CA, USA, SAGE.

PAWLINA, W., HROMANIK, M. J., MILANESE, T. R., DIERKHISING, R., VIGGIANO, T. R. & CARMICHAEL, S. W. 2006. Leadership and professionalism curriculum in the gross anatomy course. *ANNALS-ACADEMY OF MEDICINE SINGAPORE,* 35**,** 609-614.

PEARSON, K. 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London,* 58**,** 240-242.

PERERA, J., MOHAMADOU, G. & KAUR, S. 2010. The use of objective structured self- assessment and peer feedback (OSSP) for learning communication skills: Evaluation using a controlled trial. *Advances in Health Sciences Education,* 15**,** 185-193.

PHILLIPS, S. P. & CLARKE, M. 2012. More than an education: the hidden curriculum, professional attitudes and career choice. *Medical Education,* 46**,** 887-893.

PIAN-SMITH, M. C. M., SIMON, R., MINEHART, R. D., PODRAZA, M., RUDOLPH, J., WALZER, T. & RAEMER, D. 2009. Teaching residents the two-challenge rule: a simulation-based approach to improve education and patient safety. *Simulation in Healthcare,* 4**,** 84-91.

PLANT, J., CORDEN, M., MOURAD, M., O'BRIEN, B. & VAN SCHAIK, S. 2013. Understanding self- assessment as an informed process: Residents' use of external information for self-assessment of performance in simulated resuscitations. *Advances in Health Sciences Education,* 18**,** 181-192.

POLAND, B. 2002. Transcription quality. *In:* GUBRIUM, J. & HOLSTEIN, J. (eds.) *Handbook of Interview Research: Context and Method.* Thousand Oaks, CA, USA: Sage.

POLAND, B. & PEDERSON, A. 1998. Reading between the lines: Interpreting Silences in Qualitative Research. *Qualitative inquiry,* 4**,** 293-312.

POLLARD, K. C., ROSS, K. & MEANS, R. 2005. Professional issues. Nurse leadership, interprofessionalism and the modernization agenda. *British Journal of Nursing,* 14**,** 339-344.
PORATH, C. L. & EREZ, A. 2007. Does rudeness really matter? The effects of rudeness on task performance and helpfulness. *Academy of Management Journal,* 50**,** 1181-1197.
POWELL, R. A. & SINGLE, H. M. 1996. Focus groups. *International journal for quality in health care,* 8**,** 499-504.
PRESTON, C. C. & COLMAN, A. M. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica,* 104**,** 1-15.
QUALITY ASSURANCE AGENCY 2006. Code of practice for the assurance of academic quality and standards in higher education: Section 6: Assessment of students.
RALL, M. & GABA, D. M. 2005. Patient simulators. *In:* MILLER, R. (ed.) *Anesthesia.* New York: Elsevier.
RAPLEY, T. 2007. *Doing Conversation, Discourse and Document Analysis,* London, SAGE Publications.
RAUPACH, T., BROWN, J., ANDERS, S., HASENFUSS, G. & HARENDZA, S. 2013. Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. *BMC medicine,* 11**,** 61.
RAUSCH, M. J. 1997. Analyzing and Reporting Focus Group Results. *In:* KRUEGER, R. A. (ed.) *Analyzing and reporting focus group results.* Thousand Oaks, CA, USA: Sage publications.
REED, D., PRICE, E. G., WINDISH, D. M., WRIGHT, S. M., GOZU, A., HSU, E. B., BEACH, M. C., KERN, D. & BASS, E. B. 2005. Challenges in systematic reviews of educational intervention studies. *Annals of Internal Medicine,* 142**,** 1080-1089.
REES, C. E. & KNIGHT, L. V. 2007. Viewpoint: the trouble with assessing students' professionalism: theoretical insights from sociocognitive psychology. *Academic Medicine,* 82**,** 46-50.
REEVES, S., ALBERT, M., KUPER, A. & HODGES, B. D. 2008. Why use theories in qualitative research? *Bmj,* 337.
REGEHR, G., MACRAE, H., REZNICK, R. K. & SZALAY, D. 1998. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine,* 73**,** 993-7.
REITER, C., PICHERT, J. & HICKSON, G. 2012. Addressing behavior and performance issues that threaten quality and patient safety: What your attorneys want you to know. *Progress in Pediatric Cardiology,* 33**,** 37-45.
RENNIE, S. C. & CROSBY, J. R. 2002. Students' perceptions of whistle blowing: implications for self-regulation. A questionnaire and focus group survey. *Medical Education,* 36**,** 173-179.
RETHANS, J.-J., STURMANS, F., DROP, R., VAN DER VLEUTEN, C. & HOBUS, P. 1991. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *British Medical Journal,* 303**,** 1377.
REX, J. H., TURNBULL, J. E., ALLEN, S. J., VOORDE, K. V. & LUTHER, K. 2000. Systematic root cause analysis of adverse drug events in a tertiary

referral hospital. *Joint Commission Journal on Quality and Patient Safety,* 26**,** 563-575.
RITCHIE, J. & SPENCER, L. 1994. Qualitative data analysis for applied policy research. *In:* BRYMAN, A. & BURGESS, R. (eds.) *Analysing qualitative data.* London: Routledge.
ROBERTS, T. E. 2013. Assessment est mort, vive assessment. *Medical Teacher,* 35**,** 535-536.
ROBINS, L., BROCK, D. M., GALLAGHER, T., KARTIN, D., LINDHORST, T., ODEGARD, P. S., MORTON, T. H. & BELZA, B. 2008. Piloting team simulations to assess interprofessional skills. *Journal of Interprofessional Care,* 22**,** 325-328.
ROCH, S. G. & O'SULLIVAN, B. J. 2003. Frame of reference rater training issues: recall, time and behavior observation training. *International Journal of Training and Development,* 7**,** 93-107.
RODGERS, D. L., BHANJI, F. & MCKEE, B. R. 2010. Written evaluation is not a predictor for skills performance in an Advanced Cardiovascular Life Support course. *Resuscitation,* 81**,** 453-456.
RODGERS, D. L., SECURRO JR, S. & PAULEY, R. D. 2009. The effect of high-fidelity simulation on educational outcomes in an advanced cardiovascular life support course. *Simulation in Healthcare,* 4**,** 200-206.
ROFF, S. & DHERWANI, K. 2011. Development of inventory for polyprofessionalism lapses at the proto-professional stage of health professions education together with recommended responses. *Medical Teacher,* 33**,** 239-243.
ROSENBERG, D. A. & SILVER, H. K. 1984. Medical student abuse: An unnecessary and preventable cause of stress. *Journal of the American Medical Association,* 251**,** 739-742.
ROSENSTEIN, A. 2011. The quality and economic impact of disruptive behaviors on clinical outcomes of patient care. *American Journal of Medical Quality,* 26**,** 372-379.
ROSS, S., LAI, K., WALTON, J. M., KIRWAN, P. & WHITE, J. S. 2013. ‘‘I have the right to a private life’’: Medical students’ views about professionalism in a digital world. *Medical Teacher,* 35**,** 826-831.
ROWLAND-MORIN, P. A., BURCHARD, K. W., GARB, J. L. & COE, N. P. 1991. Influence of effective communication by surgery students on their oral examination scores. *Academic Medicine,* 66**,** 169-71.
ROWNTREE, D. 1987. *Assessing students: How shall we know them?*, Taylor & Francis.
ROYAL COLLEGE OF PHYSICIANS 2012. Medical workforce: New Deal and European Working Time Directive. London: Royal College of Physicians.
RUDLAND, J. & MIRES, G. 2005. Characteristics of doctors and nurses as perceived by students entering medical school: implications for shared teaching. *Medical Education,* 39**,** 448-455.
SACHS, B. P. 2005. A 38-year-old woman with fetal loss and hysterectomy. *Journal of the American Medical Association,* 294**,** 833-840.
SALAS, E., ROSEN, M. & KING, H. 2009. Integrating Teamwork into the ‘‘DNA’’ of Graduate Medical Education: Principles for Simulation-Based Training. *Journal of Graduate Medical Education,* 1**,** 243-244.

SALVATORI, P. 2001. Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education,* 6**,** 159-175.

SARGEANT, J., ARMSON, H., CHESLUK, B., DORNAN, T., EVA, K. W., HOLMBOE, E. S., LOCKYER, J., LONEY, E., MANN, K. V. & VAN DER VLEUTEN, C. 2010. The processes and dimensions of informed self-assessment: A conceptual model. *Academic Medicine,* 85**,** 1212-1220.

SATTERWHITE, R. C., SATTERWHITE, W. M. & ENARSON, C. 2000. An ethical paradox: the effect of unethical conduct on medical students' values. *Journal of medical ethics,* 26**,** 462-465.

SAVOLDELLI, G. L., NAIK, V. N., PARK, J., JOO, H. S., CHOW, R. & HAMSTRA, S. J. 2006. Value of debriefing during simulated crisis management: oral versus video-assisted oral feedback. *Anesthesiology,* 105**,** 279-285.

SCHAEFER, H. G., HELMREICH, R. L. & SCHEIDEGGER, D. 1994. Human factors and safety in emergency medicine. *Resuscitation,* 28**,** 221-225.

SCHAEFER, H. G., HELMREICH, R. L. & SCHEIDEGGER, D. 1995. Safety in the operating theatre—part 1: interpersonal relationships and team performance. *Current Anaesthesia & Critical Care,* 6**,** 48-53.

SCHILLINGER, M. 2006. *Learning environment and moral development: How university education fosters moral judgment competence in Brazil and two German-speaking countries,* Aachen, Germany, Shaker Verlag.

SCHUWIRTH, L. & ASH, J. 2013. Assessing tomorrow's learners: In competency-based education only a radically different holistic method of assessment will work. Six things we could forget. *Medical Teacher,* 35**,** 555-559.

SCHUWIRTH, L. & VAN DER VLEUTEN, C. 2004. Merging views on assessment. *Medical Education,* 38**,** 1208-1210.

SCHUWIRTH, L. & VAN DER VLEUTEN, C. P. 2011. General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher,* 33**,** 783-797.

SCHUWIRTH, L. W. T. & VAN DER VLEUTEN, C. P. M. 2003. The use of clinical simulations in assessment. *Medical Education,* 37**,** 65-71.

SCHUWIRTH, L. W. T. & VAN DER VLEUTEN, C. P. M. 2010. How to design a useful test: the principles of assessment. *In:* SWANWICK, T. (ed.) *Understanding medical education: Evidence, theory and practice.* Wiley-Blackwell.

SEALE, C. 1999. Quality in Qualitative Research. *Qualitative inquiry,* 5**,** 465-478.

SEVDALIS, N., DAVIS, R., KOUTANTJI, M., UNDRE, S., DARZI, A. & VINCENT, C. A. 2008. Reliability of a revised NOTECHS scale for use in surgical teams. *The American Journal of Surgery,* 196**,** 184-190.

SEXTON, J. B., THOMAS, E. J. & HELMREICH, R. L. 2000. Error, stress, and teamwork in medicine and aviation: cross sectional surveys. *British Medical Journal,* 320**,** 745-749.

SHAPIRO, M. J., MOREY, J. C., SMALL, S. D., LANGFORD, V., KAYLOR, C. J., JAGMINAS, L., SUNER, S., SALISBURY, M. L., SIMON, R. & JAY, G. D. 2004. Simulation based teamwork training for emergency department staff: does it improve clinical team performance when added to an existing didactic teamwork curriculum? *Quality and Safety in Health Care,* 13**,** 417-421.

SHARMA, S., BOET, S., KITTO, S. & REEVES, S. 2011. Interprofessional simulated learning: the need for'sociological fidelity'. *Journal of interprofessional care,* 25**,** 81-83.
SHEEHAN, T. J., THAL, S. E., KRAUSE, K. C., CANDEE, D., COTTON, J. & GEER, S. 1987. Improving Physician Skills in Managing Morally Problematic Cases. *Annual Meeting of the American Educational Research Association.* Washington, DC, USA.
SHROUT, P. E. & FLEISS, J. L. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin,* 86**,** 420-428.
SHUMWAY, J. M. 2004. Components of quality: competence, leadership, teamwork, continuing learning and service. *Medical Teacher,* 26**,** 397-399.
SIM, J. 1998. Collecting and analysing qualitative data: issues raised by the focus group. *Journal of Advanced Nursing,* 28**,** 345-352.
SIM, J. & SNELL, J. 1996. Focus groups in physiotherapy evaluation and research. *Physiotherapy,* 82**,** 189-198.
SINGH, H., THOMAS, E. J., PETERSEN, L. A. & STUDDERT, D. M. 2007. Medical errors involving trainees: a study of closed malpractice claims from 5 insurers. *Archives of internal medicine,* 167**,** 2030-2036.
SLAGLE, J., WEINGER, M. B., DINH, M.-T. T., BRUMER, V. V. & WILLIAMS, K. 2002. Assessment of the intrarater and interrater reliability of an established clinical task analysis methodology. *Anesthesiology,* 96**,** 1129-1139.
SLOCUMB, E. M. & COLE, F. L. 1991. A practical approach to content validation. *Applied Nursing Research,* 4**,** 192-195.
SMITH, M. W. 1995. Ethics in focus groups: a few concerns. *Qualitative health research,* 5**,** 478-486.
SMITH, R. 2009. *Doctors are the most trustworthy and journalists the least, poll finds* [Online]. Available: http://www.telegraph.co.uk/health/healthnews/4591602/Doctors-are-the-most-trustworthy-and-journalists-the-least-poll-finds.html.
SMITHSON, J. 2000. Using and analysing focus groups: limitations and possibilities. *International Journal of Social Research Methodology,* 3**,** 103-119.
SOUTHGATE, L. & VAN DER VLEUTEN, C. P. M. 2014. A conversation about the role of medical regulators. *Medical Education,* 48**,** 215-218.
SRINIVASAN, M., HWANG, J. C., WEST, D. & YELLOWLEES, P. M. 2006. Assessment of clinical skills using simulator technologies. *Academic Psychiatry,* 30**,** 505-515.
ST PIERRE, M., SCHOLLER, A., STREMBSKI, D. & BREUER, G. 2012. [Do residents and nurses communicate safety relevant concerns?: simulation study on the influence of the authority gradient]. *Der Anaesthesist,* 61**,** 857-866.
STALMEIJER, R. E., MCNAUGHTON, N. & VAN MOOK, W. N. 2014. Using focus groups in medical education research: AMEE Guide No. 91. *Medical Teacher,* 36**,** 923-939.
STASSER, G. & TITUS, W. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology,* 48**,** 1467-1478.
STERN, D. & GINSBURG, S. 2004. The Professionalism Movement: Behaviors Are the Key to Progress. *The American Journal of Bioethics,* 4**,** 14-15.

STERN, D. T. 2006. A Framework for Measuring Professionalism. *In:* STERN, D. T. (ed.) *Measuring Medical Professionalism.* Oxford: Oxford University Press.

STIRLING, K., HOGG, G., KER, J., ANDERSON, F., HANSLIP, J. & BYRNE, D. 2012. Using simulation to support doctors in difficulty. *The clinical teacher,* 9**,** 285-289.

STOLLER, J. K., ROSE, M., LEE, R., DOLGAN, C. & HOOGWERF, B. J. 2004. Teambuilding and leadership training in an internal medicine residency training program. *Journal of general internal medicine,* 19**,** 692-697.

STREINER, C. 1995. Clinical ratings—ward rating. *In:* SHANNON, S. & NORMAN, G. (eds.) *Evaluation methods: a resource handbook.* Hamilton, Ontario, Canada: Program for Educational Development, McMaster University.

SUDDABY, R. 2006. From the editors: What grounded theory is not. *Academy of management journal,* 49**,** 633-642.

SUTCLIFFE, K. M., LEWTON, E. & ROSENTHAL, M. M. 2004. Communication failures: an insidious contributor to medical mishaps. *Academic Medicine,* 79**,** 186-194.

SWANWICK, T. 2010. *Understanding medical education: Evidence, theory and practice*, John Wiley & Sons.

SWARTZ, M. H., COLLIVER, J. A., BARDES, C. L., CHARON, R., FRIED, E. D. & MOROFF, S. 1997. Validating the standardized-patient assessment administered to medical students in the New York City consortium. *Academic Medicine,* 72**,** 619-26.

SWARTZ, M. H., COLLIVER, J. A., BARDES, C. L., CHARON, R., FRIED, E. D. & MOROFF, S. 1999. Global ratings of videotaped performance versus global ratings of actions recorded on checklists: a criterion for performance assessment with standardized patients. *Academic Medicine,* 74**,** 1028-32.

SYMONDS, P. M. 1924. On the Loss of Reliability in Ratings Due to Coarseness of the Scale. *Journal of Experimental Psychology,* 7**,** 456.

TALLIS, R. C. 2006. Doctors in society: medical professionalism in a changing world. *Clinical Medicine,* 6**,** 7-12.

TEDDLIE, C. & YU, F. 2007. Mixed methods sampling a typology with examples. *Journal of mixed methods research,* 1**,** 77-100.

TEN CATE, T. & DE HAES, J. 2000. Summative assessment of medical students in the affective domain. *Medical Teacher,* 22**,** 40-43.

THE ABIM FOUNDATION, ACP-ASIM FOUNDATION & EUROPEAN FEDERATION OF INTERNAL MEDICINE 2002. Medical Professionalism in the New Millennium: A Physicians Charter. *Annals of Internal Medicine,* 136**,** 243-246.

THE JOURNAL. 2006. *Patients support GP* [Online]. The Journal. Available: http://www.thejournal.co.uk/news/north-east-news/patients-support-gp-4587721 [Accessed 11th July 2014.

THISTLETHWAITE, J. & SPENCER, J. 2008. *Professionalism in Medicine,* Abingdon, UK, Radcliffe Publishing Ltd.

THOMAS, E., SEXTON, J. & HELMREICH, R. 2004. Translating teamwork behaviours from aviation to healthcare: development of behavioural markers for neonatal resuscitation. *Quality and Safety in Health Care,* 13**,** i57-i64.

THOMAS, E. J., SEXTON, J. B. & HELMREICH, R. L. 2003. Discrepant attitudes about teamwork among critical care nurses and physicians*. *Critical care medicine,* 31**,** 956-959.

TIBERIUS, R. 2006. *The Focus Group Guide,* Miami, FL, USA, University of Miami: Miller School of Medicine.

TILL, A. D., PETTIFER, G. D., O'SULLIVAN, H. & MCKIMM, J. 2014. Developing and harnessing the leadership potential of doctors in training. *British journal of hospital medicine,* 75**,** 523-527.

TINSLEY, H. E. & WEISS, D. J. 1975. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology,* 22**,** 358.

TODRES, M., STEPHENSON, A. & JONES, R. 2007. Medical education research remains the poor relation. *BMJ: British Medical Journal,* 335**,** 333.

TWOHIG, P. L. & PUTNAM, W. 2002. Group interviews in primary care research: advancing the state of the art or ritualized research? *Family Practice,* 19**,** 278-284.

UHARI, M., KOKKONEN, J., NUUTINEN, M., VAINIONPAA, L., RANTALA, H., LAUTALA, P. & VÄYRYNEN, M. 1994. Medical student abuse: an international phenomenon. *Journal of the American Medical Association,* 271**,** 1049-1051.

UNDRE, S., KOUTANTJI, M., SEVDALIS, N., GAUTAMA, S., SELVAPATT, N., WILLIAMS, S., SAINS, P., MCCULLOCH, P., DARZI, A. & VINCENT, C. 2007a. Multidisciplinary crisis simulations: the way forward for training surgical teams. *World journal of surgery,* 31**,** 1843-1853.

UNDRE, S., SEVDALIS, N., HEALEY, A. N., DARZI, A. & VINCENT, C. A. 2007b. Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. *World journal of surgery,* 31**,** 1373-1381.

URQUHART, L. M., REES, C. E. & KER, J. S. 2014. Making sense of feedback experiences: a multi‐school study of medical students' narratives. *Medical Education,* 48**,** 189-203.

VAN DER VLEUTEN, C. & SCHUWIRTH, L. 2006. *How to Design a Useful Test: The Principles of Assessment. Understanding Medical Education.*, Association for the Study of Medical Education.

VAN DER VLEUTEN, C., SCHUWIRTH, L., SCHEELE, F., DRIESSEN, E. & HODGES, B. 2010. The assessment of professional competence: building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology,* 24**,** 703-719.

VAN DER VLEUTEN, C. P. 2000. Validity of final examinations in undergraduate medical training. *British Medical Journal,* 321**,** 1217.

VAN DER VLEUTEN, C. P. 2012. Towards a systems approach to assessment. *Medical Teacher,* 34**,** 185-186.

VAN DER VLEUTEN, C. P., NORMAN, G. R. & GRAAFF, E. 1991. Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education,* 25**,** 110-118.

VAN DER VLEUTEN, C. P. & SCHUWIRTH, L. 2010. How to Design a Useful Test: The Principles of Assessment. *In:* SWANWICK, T. (ed.) *Understanding Medical Education.*: Association for the Study of Medical Education.

VAN DER VLEUTEN, C. P. & SCHUWIRTH, L. W. 2005. Assessing professional competence: from methods to programmes. *Medical Education,* 39**,** 309-317.

VAN MOOK, W., VAN LUIJK, S., O'SULLIVAN, H., WASS, V., SCHUWIRTH, L. & VAN DER VLEUTEN, C. 2009a. General considerations regarding assessment of professional behaviour. *European Journal of Internal Medicine,* 20**,** e90-e95.

VAN MOOK, W. N. K. A., GORTER, S. L., DE GRAVE, W. S., VAN LUIJK, S. J., O'SULLIVAN, H., WASS, V., ZWAVELING, J. H., SCHUWIRTH, L. W. & VAN DER VLEUTEN, C. P. M. 2009b. Professionalism beyond medical school: an educational continuum? *European journal of internal medicine,* 20**,** e148-e152.

VAN MOOK, W. N. K. A., GORTER, S. L., KIEBOOM, W., CASTERMANS, M. G. T. H., DE FEIJTER, J., DE GRAVE, W. S., ZWAVELING, J. H., SCHUWIRTH, L. W. T. & VAN DER VLEUTEN, C. P. M. 2012. Poor professionalism identified through investigation of unsolicited healthcare complaints. *Postgraduate medical journal,* 88**,** 443-450.

VAN MOOK, W. N. K. A., GORTER, S. L., O'SULLIVAN, H., WASS, V., SCHUWIRTH, L. W. & VAN DER VLEUTEN, C. P. M. 2009c. Approaches to professional behaviour assessment: Tools in the professionalism toolbox. *European journal of internal medicine,* 20**,** e153-e157.

VAN MOOK, W. N. K. A., VAN LUIJK, S. J., DE GRAVE, W., O'SULLIVAN, H., WASS, V., SCHUWIRTH, L. W. & VAN DER VLEUTEN, C. P. M. 2009d. Teaching and learning professional behavior in practice. *European journal of internal medicine,* 20**,** e105-e111.

VAN ROOYEN, M. 2004. The views of medical students on professionalism in South Africa. *South African Family Practice,* 46**,** 29-32.

VARKEY, P., PELOQUIN, J., REED, D., LINDOR, K. & HARRIS, I. 2009. Leadership curriculum in undergraduate medical education: A study of student and faculty perspectives. *Medical Teacher,* 31**,** 244-250.

VLEUTEN, C. P. M., LUYK, S. J. & BECKERS, H. J. M. 1989. A written test as an alternative to performance testing. *Medical Education,* 23**,** 97-107.

WAGNER, D. P., HOPPE, R. B. & LEE, C. P. 2009. The patient safety OSCE for PGY-1 residents: a centralized response to the challenge of culture change. *Teaching and learning in medicine,* 21**,** 8-14.

WALLIN, C.-J., MEURLING, L., HEDMAN, L., HEDEGARD, J. & FELLANDER-TSAI, L. 2007. Target-focused medical emergency team training using a human patient simulator: effects on behaviour and attitude. *Medical Education,* 41**,** 173-180.

WALTON, M. M. 2006. Hierarchies: the Berlin Wall of patient safety. *Quality and Safety in Health Care,* 15**,** 229-230.

WANG-CHENG, R. M., FULKERSON, P. K., BARNAS, G. P. & LAWRENCE, S. L. 1995. Effect of student and preceptor gender on clinical grades in an ambulatory care clerkship. *Academic Medicine,* 70**,** 324-6.

WATMOUGH, S., GARDEN, A. & TAYLOR, D. 2006a. Does a new integrated PBL curriculum with specific communication skills classes produce Pre Registration House Officers (PRHOs) with improved communication skills? *Medical Teacher,* 28**,** 264-269.

WATMOUGH, S., GARDEN, A. & TAYLOR, D. 2006b. Pre-registration house officers' views on studying under a reformed medical curriculum in the UK. *Medical Education,* 40**,** 893-899.

WATTS, M. & EBBUTT, D. 1987. More than the sum of the parts: research methods in group interviewing. *British Educational Research Journal,* 13**,** 25-34.

WEAR, D. & KUCZEWSKI, M. G. 2004. The professionalism movement: Can we pause? *American Journal of Bioethics,* 4**,** 1-10.

WELLER, J., SHULRUF, B., TORRIE, J., FRENGLEY, R., BOYD, M., PAUL, A., YEE, B. & DZENDROWSKYJ, P. 2013. Validation of a measurement tool for self-assessment of teamwork in intensive care. *British journal of anaesthesia,* 111**,** 460-467.

WELLER, J. M. 2004. Simulation in undergraduate medical education: bridging the gap between theory and practice. *Medical Education,* 38**,** 32-38.

WHITEHEAD, C. R., HODGES, B. D. & AUSTIN, Z. 2013. Dissecting the doctor: from character to characteristics in North American medical education. *Advances in Health Sciences Education,* 18**,** 687-699.

WHITTINGHAM, R. B. 2004. *The Blame Machine: Why Human Error Causes Accidents,* Oxford, Elsevier Butterworth-Heinemann.

WILKINSON, T. J. & FRAMPTON, C. M. 2003. Assessing performance in final year medical students. Can a postgraduate measure be used in an undergraduate setting? *Medical Education,* 37**,** 233-240.

WILKINSON, T. J., WADE, W. B. & KNOCK, L. 2009. A blueprint to assess professionalism – Results of a systematic review. *Academic Medicine,* 84**,** 551-558.

WONG, A. & TROLLOPE‐KUMAR, K. 2014. Reflections: an inquiry into medical students' professional identity formation. *Medical Education,* 48**,** 489-501.

WRIGHT, M. C., PHILLIPS-BUTE, B. G., PETRUSA, E. R., GRIFFIN, K. L., HOBBS, G. W. & TAEKMAN, J. M. 2009. Assessing teamwork in medical education and practice: relating behavioural teamwork ratings and clinical performance. *Medical Teacher,* 31**,** 30-38.

XYRICHIS, A. & LOWTON, K. 2008. What fosters or prevents interprofessional teamworking in primary and community care? A literature review. *International journal of nursing studies,* 45**,** 140-153.

YEE, B., NAIK, V. N., JOO, H. S., SAVOLDELLI, G. L., CHUNG, D. Y., HOUSTON, P. L., KARATZOGLOU, B. J. & HAMSTRA, S. J. 2005. Nontechnical skills in anesthesia crisis management with repeated exposure to simulation-based education. *Anesthesiology,* 103**,** 241-248.

YOUNG, J. S., DUBOSE, J. E., HEDRICK, T. L., CONAWAY, M. R. & NOLLEY, B. 2007. The use of "war games" to evaluate performance of students and residents in basic clinical scenarios: a disturbing analysis. *Journal of Trauma-Injury, Infection, and Critical Care,* 63**,** 556-564.

YOUNGBLOOD, P., HARTER, P. M., SRIVASTAVA, S., MOFFETT, S., HEINRICHS, W. L. & DEV, P. 2008. Design, development, and evaluation of an online virtual emergency department for training trauma teams. *Simulation in Healthcare,* 3**,** 146-153.

YULE, S., FLIN, R., MARAN, N., ROWLEY, D., YOUNGSON, G. & PATERSON-BROWN, S. 2008. Surgeons' non-technical skills in the operating room:

reliability testing of the NOTSS behavior rating system. *World journal of surgery,* 32**,** 548-556.

YULE, S., FLIN, R., PATERSON-BROWN, S., MARAN, N. & ROWLEY, D. 2006. Development of a rating system for surgeons' non-technical skills. *Medical Education,* 40**,** 1098-1104.

YULE, S., ROWLEY, D., FLIN, R., MARAN, N., YOUNGSON, G., DUNCAN, J. & PATERSON-BROWN, S. 2009. Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. *ANZ journal of surgery,* 79**,** 154-160.

APPENDICES

Appendix 2-1: Data extraction form

Reference:	
Country	

Does the article describe a tool for assessing teamwork and/or leadership?	YES/NO (If NO then reject from further review; If YES then a) Teamwork, b) Leadership or c) Both)
Is the tool described for use on an individual (not a team)	YES/NO (If NO then reject from further review after completing the teamwork and leadership subheadings below)
Is the tool described for use in healthcare?	YES/NO (If NO then reject from further review)

Study design (e.g. RCT, cohort, survey, pre/post, pilot)	
Primary aim of study? (Evaluation of current tool, design and evaluation of new tool, other)	
Was it a single intervention (e.g. scenario, OSCE) or longer-term?	
Who was being assessed? (e.g. medical student, junior doctors, surgeons)	
UG Nurses/Medics/Other	
PG Nurses/Medics/Other	
How were the participants recruited? (e.g. random, purposeful, convenience, other, not specified)	
Who was assessing? (e.g. medical faculty, consultants, psychologists)	
Where did the assessment take place? (e.g. workplace, simulator)	

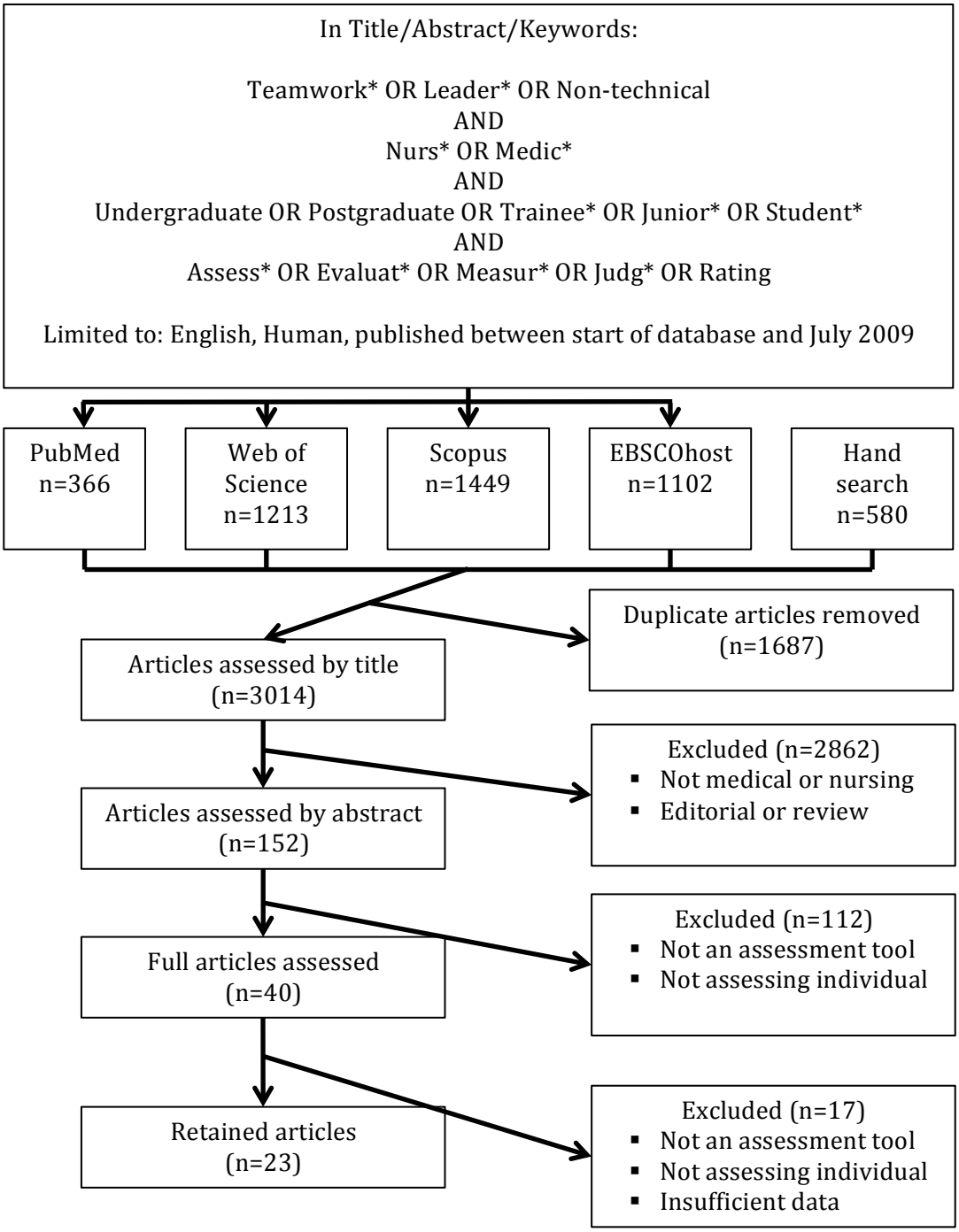
How was the tool developed?	
How was the assessment carried out?	
Evidence of validity of tool?	YES/NO (If YES then what type, e.g. construct, content, criterion)
Evidence of reliability of tool?	YES/NO (If YES then what type, e.g. inter-rater, test-retest)
Evidence of feasibility of tool?	YES/NO (If YES then what?)
Evidence of educational impact?	YES/NO (If YES then Kirkpatrick Level*: 1/2/3/4)
Evidence of acceptability?	YES/NO (If YES then what?)

TEAMWORK SUBHEADINGS:

LEADERSHIP SUBHEADINGS:

Additional comments:

Appendix 2-2: Search strategy results



Appendix 2-3: Final references used in analysis

1) BRETT-FLEEGLER, M. B., VINCI, R. J., WEINER, D. L., HARRIS, S. K., SHIH, M.-C. & KLEINMAN, M. E. 2008. A simulator-based tool that assesses pediatric resident resuscitation competency. *Pediatrics,* 121**,** e597-e603.
2) BRINKMAN, W. B., GERAGHTY, S. R., LANPHEAR, B. P., KHOURY, J. C., DEL REY, J. A. G., DEWITT, T. G. & BRITTO, M. T. 2006. Evaluation of resident communication skills and professionalism: A matter of perspective? *Pediatrics,* 118**,** 1371-1379.
3) CHRISTENSON, J., PARRISH, K., BARABE, S., NOSEWORTHY, R., WILLIAMS, T., GEDDES, R. & CHALMERS, A. 1998. A comparison of multimedia and standard advanced cardiac life support learning. *Academic emergency medicine,* 5**,** 702-708.
4) COOPER, S. & WAKELAM, A. 1999. Leadership of resuscitation teams: Lighthouse Leadership. *Resuscitation,* 42**,** 27-45.
5) DANNEFER, E. F., HENSON, L. C., BIERER, S. B., GRADY-WELIKY, T. A., MELDRUM, S., NOFZIGER, A. C., BARCLAY, C. & EPSTEIN, R. M. 2005. Peer assessment of professional competence. *Medical Education,* 39**,** 713-722.
6) DIMATTEO, M. R. & DINICOLA, D. D. 1981. Sources of assessment of physician performance: a study of comparative reliability and patterns of intercorrelation. *Medical care***,** 829-842.
7) EPSTEIN, R. M., DANNEFER, E. F., NOFZIGER, A. C., HANSEN, J. T., SCHULTZ, S. H., JOSPE, N., CONNARD, L. W., MELDRUM, S. C. & HENSON, L. C. 2004. Comprehensive assessment of professional competence: the Rochester experiment. *Teaching and learning in medicine,* 16**,** 186-196.
8) FLETCHER, G., FLIN, R., MCGEORGE, P., GLAVIN, R., MARAN, N. & PATEY, R. 2003. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *British Journal of Anaesthesia,* 90**,** 580-588.

9) GILFOYLE, E., GOTTESMAN, R. & RAZACK, S. 2007. Development of a leadership skills workshop in paediatric advanced resuscitation. *Medical teacher,* 29**,** e276-e283.
10) KAYE, W. & MANCINI, M. E. 1986. Use of the Mega Code to evaluate team leader performance during advanced cardiac life support. *Critical care medicine,* 14**,** 99-104.
11) KIM, J., NEILIPOVITZ, D., CARDINAL, P., CHIU, M. & CLINCH, J. 2006. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Critical care medicine,* 34**,** 2167-2174.
12) MISHRA, A., CATCHPOLE, K., DALE, T. & MCCULLOCH, P. 2008. The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surgical endoscopy,* 22**,** 68-73.
13) MOORTHY, K., MUNZ, Y., ADAMS, S., PANDEY, V. & DARZI, A. 2005. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. *Annals of surgery,* 242**,** 631-639.
14) MOORTHY, K., MUNZ, Y., FORREST, D., PANDEY, V., UNDRE, S., VINCENT, C. & DARZI, A. 2006. Surgical crisis management skills training and assessment: a stimulation-based approach to enhancing operating room performance. *Annals of surgery,* 244**,** 139-147.
15) MORISON, S. L. & STEWART, M. C. 2005. Developing interprofessional assessment. *Learning in Health & Social Care,* 4**,** 192-202.
16) ORLANDER, J. D., WIPF, J. E. & LEW, R. A. 2006. Development of a tool to assess the team leadership skills of medical residents. *Medical Education Online,* 11**,** 1-6.
17) PAWLINA, W., HROMANIK, M. J., MILANESE, T. R., DIERKHISING, R., VIGGIANO, T. R. & CARMICHAEL, S. W. 2006. Leadership and professionalism curriculum in the gross anatomy course. *ANNALS-ACADEMY OF MEDICINE SINGAPORE,* 35**,** 609-614.
18) ROBINS, L., BROCK, D. M., GALLAGHER, T., KARTIN, D., LINDHORST, T., ODEGARD, P. S., MORTON, T. H. & BELZA, B. 2008. Piloting team

simulations to assess interprofessional skills. *Journal of Interprofessional Care,* 22**,** 325-328.

19) SEVDALIS, N., DAVIS, R., KOUTANTJI, M., UNDRE, S., DARZI, A. & VINCENT, C. A. 2008. Reliability of a revised NOTECHS scale for use in surgical teams. *The American Journal of Surgery,* 196**,** 184-190.
20) WILKINSON, T. J. & FRAMPTON, C. M. 2003. Assessing performance in final year medical students. Can a postgraduate measure be used in an undergraduate setting? *Medical Education,* 37**,** 233-240.
21) WRIGHT, M. C., PHILLIPS-BUTE, B. G., PETRUSA, E. R., GRIFFIN, K. L., HOBBS, G. W. & TAEKMAN, J. M. 2009. Assessing teamwork in medical education and practice: relating behavioural teamwork ratings and clinical performance. *Medical teacher,* 31**,** 30-38.
22) YOUNGBLOOD, P., HARTER, P. M., SRIVASTAVA, S., MOFFETT, S., HEINRICHS, W. L. & DEV, P. 2008. Design, development, and evaluation of an online virtual emergency department for training trauma teams. *Simulation in Healthcare,* 3**,** 146-153.
23) YULE, S., FLIN, R., MARAN, N., ROWLEY, D., YOUNGSON, G. & PATERSON-BROWN, S. 2008. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World journal of surgery,* 32**,** 548-556.

APPENDIX 2-4: Sample size, graduation status and profession

Study number	Sample size	Undergraduate (UG)/Postgraduate (PG)/Both (Nurses/Medics/ Other)
1	25	PG (0/25/0)
2	36	PG (0/36/0)
3	113	UG (0/113/0)
4	18	PG (0/18/0)
5	97	UG (0/97/0)
6	40	PG (0/40/0)
7	95	UG (0/95/0)
8	8 (scripted)	PG (0/8/0)
9	20	PG (0/20/0)
10	41	PG (9/32/0)
11	59	PG (0/59/0)
12	Not specified	PG
13	27	PG (0/27/0)
14	20	PG (0/20/0)
15	31	UG (12/19/0)
16	134	PG (0/134/0)
17	43	UG (0/43/0)
18	15	UG (5/5/5)
19	Not specified	PG
20	123	UG (0/123/0)
21	35	UG (0/35/0)
22	30	Both (0/30/0)
23	6 (scripted)	PG (0/6/0)

APPENDIX 2-5: STUDY DESIGNS

Study number	Study Design	Simulated (S) or Real (R)	Short-term (S) or Long-term (L)	Assessors (Peer/Non-Peer)
1	Pilot	S	S	N
2	Other	R	L	N
3	Pre/post (blinded)	S	S	N
4	Observational	R	S	N
5	Survey	R	L	P
6	Pilot	R	L	N
7	Survey	R	L	P
8	Pilot	S	S	P
9	Pre/post	S	S	?
10	Pilot	S	S	?
11	Pilot	S	S	N
12	Pilot	R	S	P
13	Pilot	S	S	N
14	Pilot	S	S	N
15	Pilot	S	S	N
16	Survey	R	L	N
17	Pilot	R	L	P
18	Pilot	S	S	N
19	Pilot	S	S	N
20	Survey	R	L	N
21	Pilot	S	S	N
22	Pre/post	S	S	N
23	Pilot	S	S	P

APPENDIX 2-6: VALIDITY

Study number	Type(s) of validity	Evidence
1	Content Construct	Content: Modified Delphi technique with panel of 13 experts to come up with 72-item tool Construct: Could the tool differentiate between trainees at different levels? Trend but not statistically significant difference in leadership scores for more advanced trainees.
2	Content Construct	Content: adaptations from current instruments, explained how raters trained; Construct: Internal consistency.
3	Content	Content: According to the authors it is "it is simple and intuitive and has face validity" (p.707)
4	Construct	Construct: Some correlation between leadership and team dynamics
5	Content Criterion	Criterion: Correlation with mock exam, final grade, and some other scores (SP rating of communication, computer exercise score) Content: Poor content validity: terms were selected based on those behaviours most likely to be consistently observed by peers in the medical school environment
6	Construct	Construct: Correlation between different feedback scores (patient, peer, attending, self)
7	Construct	Construct: Correlation between peer assessment, SP evaluation, and Rochester Communication Rating Scale
8	Content	Content: Completeness: (by questionnaire of the consultants: Did it address the key behaviours displayed? Anything missing? Anything superfluous?) Observability: 13 elements

		observable >80% and all categories observable >95%
9	Content	Content: "Our checklist was not formally validated before it was used. However, we derived our checklists from a previously validated Crisis Resource Management curriculum (Gaba et al. 1998), as well as a well-recognized standard for education of resuscitation skills in paediatrics, the Pediatric Advanced Life Support (PALS) course (American Heart Association 2001). Further work needs to be done to thoroughly validate our checklist. p.e279"
10	Construct	Construct: Negatively: the more experienced physicians performed more poorly
11	Construct Content	Content: Traits from Ottawa GRS follow those set out by Gaba and were reviewed by simulation and CRM instructors across Canada; trained support staff and raters, residents received orientation, identical scenarios Construct: able to discriminate between PGY-1/-3; PGY-3 performance better than PGY-1
12	Construct	Construct: The worse the situational awareness score the more technical errors made.
13	Construct	Construct: Differences between the junior and middle level trainees in leadership
14	Construct Content	Construct: No difference in human factors skills between junior and senior trainees Content: still needs to be done using a task analysis (Fletcher) or Delphi-type questionnaire; Face validity of the simulation, rather than the tool
15	None	There is a mention of validity and "data

		triangulation" but this is not referred to again within the article.
16	Construct Content	Construct: Correlated the scores from their assessment with a validated measure of teaching skills (Clinical Teaching Assessment Form) and with a global rating score from the residency program director. Used Pearson correlation coefficient. (0.45 between RLS and PD, 0.87 between RLS and CTAF, 0.9 between mean of 6-item RLS and 7th item (global rating)). Content: The authors say that the RLS "is short, simply stated, and has face validity" (p.5)
17	Construct	Construct: Some correlation with average team exam scores
18	None	
19	Content	Content: Discussion around development of tool and reasoning behind addition of communication into scale.
20	Construct	Construct: Correlating ratings with traditional assessments. High correlations between different types of raters
21	Construct	Construct: Some correlation between teamwork and team performance.
22	Criterion	Criterion: Able to identify improved performance over time
23	None	

APPENDIX 2-7: RELIABILITY

Study number	Type of reliability	Evidence	Scores
1	Calibration Inter-rater Percentage agreement	Scoring was standardised by a faculty development session with 2 tool developers and 2 physician expert raters. Only one videotape was scored to standardise. Inter-rater reliability using intraclass correlation correlation coefficients (ICC) (for domain and summary scores only) and Cohen's K for individuals but problems with this mean that they also gave percentage agreements.	ICC for Leadership: 0.74 Cohen's K: 0 or –ve for 19 items (do not specify if this is "leadership" or not) 0.51 for the remainder % of exact agreement for leadership: 85.1%
2	Inter-rater Internal consistency	A suggestion (with no data) that parents and attending physicians rated similar. Internal consistency (internal reliability) (Cronbach's alpha was high)	Cronbach's alpha: nurse evaluation 0.96, attending evaluation 0.91
3	None	"This scoring system was thought to be intuitive. It has not been tested formally for reliability. "We intend to formally test the inter- and intra-rater reliability in a future study." No rater training.	N/A

4	Inter-rater	2nd rater only scored 2 videos. Cohen's kappa.	Cohen's kappa: 0.72 and 0.71
5	Inter-rater	Cronbach's alpha. No inter-rater score but instead data on variability and number of raters needed (approx. 6) in order to achieve generalizability coefficient of 0.7.	Cronbach's alpha (for work habits): 0.94 Generalisability co-efficient: 0.7
6	Internal consistency	Cronbach's alpha	Cronbach's alpha (interpersonal skill): 0.90
7	None	N/A	N/A
8	Calibration Inter-rater agreement Internal consistency	4hrs of training for rater standardisation. Internal consistency (Cronbach's alpha). Inter-rater r_{wg}	Cronbach's alpha: 0.79-0.86 r_{wg} Team-working elements: 0.58-0.66; Team-working category: 0.65
9	None	N/A	N/A
10	None		
11	Inter-rater Internal consistency Intra-rater	Inter-rater and Intra-rater reliability and internal consistency: Internal consistency (Cronbach's alpha) and interrater reliability (ICC). Intra-rater (provide means, mean differences and p value but do not specify test)	Cronbach's alpha: not provided in text ICC (Leadership): 0.491 and 0.626
12	Calibration	10 of the lap	Cronbach's alpha:

	Inter-rater	cholecystectomies had 2nd rater: Cronbach's alpha Clinical research fellow was trained by a retired pilot (TD) through a process which continued until their independent scores were in good agreement (p.69) but no indication of how long this took.	0.88 (provided for total team score only and not for individual items)
13	Calibration Inter-rater	Inter-rater: First 5 rated together.	Cronbach's alpha: 0.84 (with 13 elements and 5-point Likert)
14	Calibration Inter-rater Internal consistency	Calibration by looking at first five videos together. Inter-rater reliability by looking at the "Intraclass efficient" Cronbach's alpha for internal consistency	ICC (for non-technical skills): 0.87 Cronbach's alpha: 0.87
15	None	There is no data about scores given except for this (p.197): The overall score obtained by students ranged from 53% to 82% (median: 67%) and there was no observable difference between the range of scores of medical and nursing students.	N/A
16	Internal consistency	Mention of Cronbach's alpha and a suggestion that the individual scores correlate well with one another but no	Cronbach's alpha: 0.98

		discussion of inter-rater reliability. “Given its use in only one institution and only on one medical service, the generalizability of the results may be limited. (p.5)” Scores from interns were generally high and had narrow ranges, suggesting a “need for training in evaluation. (p.5)” The ratio of items to sample size of our pilot analysis was suboptimal, potentially impacting the reliability of the instrument. (p.5)”	
17	None		N/A
18	None		N/A
19	Internal consistency	Cronbach’s alpha. No discussion of how raters were trained.	Cronbach’s alpha (Leadership and managerial skills): 0.81, 0.87
20	Internal consistency	Internal consistency across the 12 items (Cronbach’s alpha) 0.973	Cronbach’s alpha: 0.973
21	Inter-rater	Inter-rater using Pearson correlations (why?) for each of the four cases and the total sums (not individual ratings). No indication of how the behavioural scientists were trained. No reliability measures for the	Pearson correlations: Only moderate agreement (0.47,0.58,0.58,0.73) between raters.

		checklist part of the study (looking at "clinical" team performance)	
22	Inter-rater Internal consistency	Inter-rater reliability with ICC internal consistency measured using Cronbach's alpha.	ICC: 0.71 Cronbach's alpha: 0.96
23	Calibration Inter-rater agreement	3 videos selected for pre-experiment training. Need more in-depth training and calibration of raters. r_{wg} across experimental groups. ICC single and average	r_{wg} (Leadership: 0.72; Communication & Teamwork: 0.7) ICC (single) (Leadership: 0.66; Communication & Teamwork: 0.63) ICC (average) (Leadership: 0.99; Communication & Teamwork: 0.99)

APPENDIX 2-8: SCOPUS FIRST 25 REJECTED ARTICLES

Authors	Title	Year	Source
Lebbon, C., Davies, S., Shippen, J.	User-centred research methods in postgraduate teaching	2009	DS 59: Proceedings of E and PDE 2009, the 11th Engineering and Product Design Education Conference - Creating a Better World
Makhdoom, N.M.	Assessment of the quality of educational climate during undergraduate clinical teaching years in the College of Medicine, Taibah University	2009	Journal of Taibah University Medical Sciences
Dalal, M., Skeete, R., Yeo, H.L., Lucas, G.I., Rosenthal, M.S.	A Physician Team's Experiences in Community-Based Participatory Research. Insights into Effective Group Collaborations20	2009	American Journal of Preventive Medicine
Shulman, K.I., Fischer, H.D., Herrmann, N., Huo, C.Y., Anderson, G.M., Rochon, P.A.	Current prescription patterns and safety profile of irreversible monoamine oxidase inhibitors: A population-based cohort study of older adults	2009	Journal of Clinical Psychiatry
Zakariasen, K.	Public health leadership: Building a graduate program and a culture	2009	International Journal of Learning
Sattenstall, M., Freeman, S.	Integrated learning: An EBL approach to pharmaceutical chemistry	2009	Pharmacy Education
Hannah, S., McConnell, J.	Serratia marcescens: A case history to illustrate the value of radiographer history taking in the face of poor health professional communication	2009	Radiography
O'Leary, K.J., Wayne, D.B., Landler, M.P., Kulkarni, N., Haviley, C., Hahn, K.J., Jeon, J., Englert, K.M., Williams, M.V.	Impact of localizing physicians to hospital units on nurse-physician communication and agreement on the plan of care	2009	Journal of General Internal Medicine
Anderson, E.S.,	The Leicester model of interprofessional	2009	Journal of Interprofessional Care

Lennox, A.	education: Developing, delivering and learning from student voices for 10 years		
Kenward, L., Stiles, M.	Intermediate care: An interprofessional education opportunity in primary care	2009	Journal of Interprofessional Care
Ross, A., Reid, S.	The retention of community service officers for an additional year at district hospitals in KwaZulu-Natal and the Eastern Cape and Limpopo provinces	2009	South African Family Practice
Hochstein, D., Moses, S., Jones, D.	Expanding your horizons : A STEM career conference for 7th and 8th grade girls	2009	ASEE Annual Conference and Exposition, Conference Proceedings
Martimianakis, M.A., McNaughton, N., Tait, G.R., Waddell, A.E., Lieff, S., Silver, I., Hodges, B.	The research innovation and scholarship in education program: An innovative way to nurture education	2009	Academic Psychiatry
Lorimer, J., Hilliard, A.	Incorporating learning technologies into undergraduate radiography education	2009	Radiography
Clark, P.G.	Reflecting on reflection in interprofessional education: Implications for theory and practice	2009	Journal of Interprofessional Care
Henry, K.J., Van Lunen, B.L., Udermann, B., OÃ±ate, J.A.	Curricular satisfaction levels of national athletic trainers' association-accredited postprofessional athletic training graduates	2009	Journal of Athletic Training
Doarn, C.R., Latifi, R., Hadeed, G., Haxhihamza, K., Bekteshi, F., Lecaj, I.	Third intensive balkan telemedicine and e-health seminar: Current principles and practices of telemedicine and e-health-clinical applications and evidence-based outcomes: International conference on telemedicine and e-health february 6-7, 2009 Skopje, Macedonia	2009	Telemedicine and e-Health
Levine, R.S., Connor, A.M., Feltbower, R.G., Robinson, M., Rudolf, M.C.J.	Weighing and measuring primary school children: Evaluation of the TRENDS model for implementation of department of health guidelines	2009	Child: Care, Health and Development
Gould, E., Reed, P.	Alzheimer's association quality care campaign and professional training initiatives: Improving hands-on care for people with dementia in the	2009	International Psychogeriatrics

	U.S.A.		
Terenius, L.	At crossroads between laboratory disciplines and medical advancements-The center for molecular medicine at the karolinska university hospital	2009	Journal of Molecular Medicine
Dyrbye, L., Cumyn, A., Day, H., Heflin, M.	A qualitative study of physicians' experiences with online learning in a masters degree program: Benefits, challenges, and proposed solutions	2009	Medical Teacher
Rosen, J.M., Long, S.A., McGrath, D.M., Greer, S.E.	Simulation in plastic surgery training and education: The path forward	2009	Plastic and Reconstructive Surgery
Westberg, J.	Making a difference: An interview with Sarah Kiguli	2008	Education for Health: Change in Learning and Practice
Omer, L., O'Sullivan, P., Masters, S., Souza, K., TachÃ©, S., Hickson, G., Mkony, C., Kaaya, E., Loeser, H.	Collaboration between academic institutions towards faculty development for educators	2008	Education for Health: Change in Learning and Practice
Baker, T.B., McFall, R.M., Shoham, V.	Current status and future prospects of clinical psychology: Toward a scientifically principled approach to mental and behavioral health care	2008	Psychological Science in the Public Interest, Supplement

Appendix 3-1: References used to develop focus group questions

Focus group methodology: Question-setting

1) ASBURY, J.-E. 1995. Overview of focus group research. *Qualitative health research,* 5**,** 414-420.
2) CAREY, M. A. 1995. Comment: concerns in the analysis of focus group data. *Qualitative health research,* 5**,** 487-495.
3) CÔTÉ-ARSENAULT, D. & MORRISON-BEEDY, D. 1999. Practical advice for planning and conducting focus groups. *Nursing Research,* 48**,** 280-283.
4) KITZINGER, J. 1995. Introducing focus groups. *British Medical Journal,* 311**,** 299-302.
5) MORGAN, D. L. 1995. Why things (sometimes) go wrong in focus groups. *Qualitative health research,* 5**,** 516-523.
6) MORRISON-BEEDY, D., CÔTÉ-ARSENAULT, D. & FEINSTEIN, N. F. 2001. Maximizing results with focus groups: Moderator and analysis issues. *Applied Nursing Research,* 14**,** 48-53.

Articles using focus group methodology providing sample questions

1) ELWYN, G., EDWARDS, A., GWYN, R. & GROL, R. 1999. Towards a feasible model for shared decision making: focus group study with general practice registrars. *Bmj,* 319**,** 753-756
2) KLABER, R. E., ROUECHE, A., HODGKINSON, R. & DAWN CASS, H. 2008. A structured approach to planning a work-based leadership development programme for doctors in training. *The International Journal of Clinical Leadership,* 16**,** 121-129.
3) SAIDI, G. & WEINDLING, A. M. 2003. An evaluation of a national scheme for continuing professional development (CPD) for career grade doctors: the Royal College of Paediatrics and Child Health's programme for paediatricians evaluated by focus group methodology. *Medical education,* 37**,** 328-334.

4) STOLLER, J. K., ROSE, M., LEE, R., DOLGAN, C. & HOOGWERF, B. J. 2004. Teambuilding and leadership training in an internal medicine residency training program. *Journal of general internal medicine,* 19**,** 692-697.

Appendix 3-2: Transcription notation and examples from code book

Transcription notation adapted from Poland (2002)

…	Pause
((coughs))	
((sighs))	
((sneezes))	
((laughs))	
((laughing))	One person
((laughter))	Several people
[	At beginning of over-lapping speech
[read ? said?]	Word unclear
()	Unable to decipher.
EMPHASIS	
Ver-y-y-y-y-y	Held sound
XXX	Name of another participant

Code	Definition	Example
Unprofessional behaviour vs. free speech	When participants refer to free speech or similar examples of being allowed to speak one's mind	I think the the sad thing about it I think is that like it's like your freedom of like speech It is important to chat you know talk to medical students about cases that you've seen
Doctor as a human	When participants refer to humans, human weakness, e.g. as opposed to an ideal state	I'm a real person, this is my job at the end of the day and I can have a laugh about it But the third one is you know

		part of their human... part of their personality
Lives in their hands	When participants use this symbolic language referring to the power of doctors	It's kinda the ultimate isn't it really with a doctor you're puttin' people put their lives in your hands. "cos like the doctor's got more patients in their like lives in their hands than the medical student at that present time.
Medicine as a job	When participants refer to Medicine as a job rather than a vocation/profession	But like you know you're not, you're there to do to do a job first and foremost. And it and it's like you know it is really just a job.
Rumour mill	When participants refer to hearsay or rumours or things that happened to friends of friends.	"You hear stories about boys about students getting pulled up for [things But there are a number of stories of what some medical students have managed to get away with

Appendix 3-3: Ethical approval

School of Medical Education

Faculty of Medicine
Cedar House
Ashton Street
Liverpool
L69 3GE

0151 795 4356
0151 794 8763
www.liv.ac.uk

Dear
T
F
W

Dr. Arpan Guha
Honorary Senior Lecturer
School of Medical Education
Cedar House
Ashton Street
Liverpool
L69 3GE

30 October 2008

Dr. Guha

Re: Ethics approval for study 200810032.

Examining professionalism in medical undergraduates.

I am pleased to inform you that the School of Medical Education Research Sub-group has given ethical approval for the above study.

Kind regards

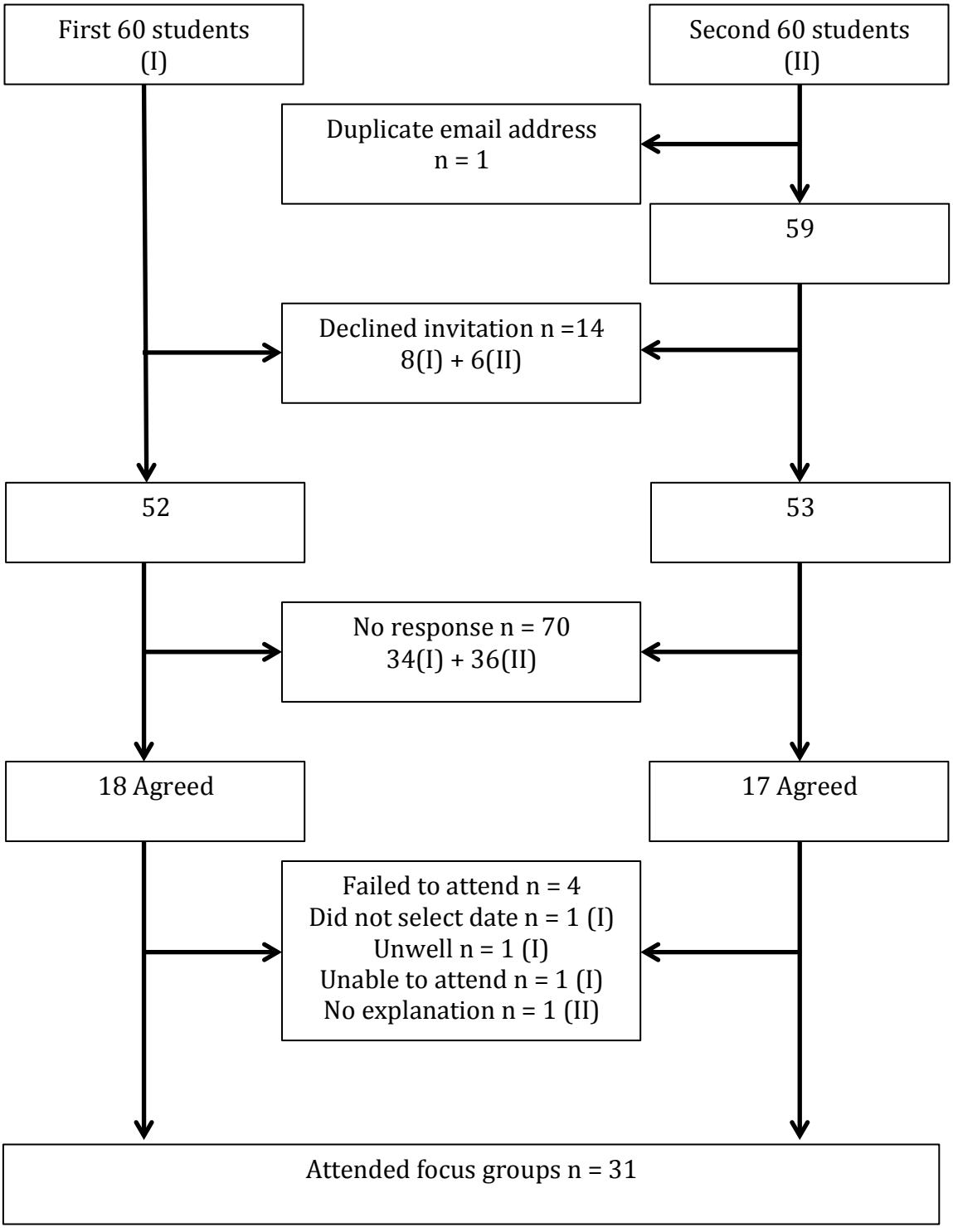
Louise Jaeger

Research Sub-group secretary.

E: Jaegerl@liverpool.ac.uk

T: 0151 795 4356

Appendix 3-4: Selection process



Appendix 3-5: Focus group questions

1) Opening question
    a. I would like everybody to tell us what their favourite bit of 4th year has been so far and why
2) Introductory question
    a. What is the first thing that comes to mind when you hear the word "professionalism"?
3) Transition question
    a. Is the professionalism expected of medical students different from that expected of doctors and, if so, why?
4) Transition question (Added after FG2. Touched on in FG2 but felt that not sufficiently explored, therefore formalised into question)
    a. If you see unprofessional behaviour by another medical student, how do you deal with that?
5) Transition question
    a. Do you think that what we think of as "professionalism" today is different from what people would have thought of as "professionalism" 30, 40 or 50 years ago and, if so, why?
6) Key questions: Professionalism
    a. Think of somebody you've met or seen at work who you think is "professional" what did they do or say, how did they act, to make you think this of them?
    b. What do you think about "bringing the profession into disrepute"? Is that still relevant today?
7) Key question: Teamwork/leadership
    a. One of the elements mentioned before was teamwork. Can you think of a really good team of people that you've seen work together and tell me what did the people in that team do to make it work so well?
    b. Can you think of a team of people that you've seen where the team didn't work very well? What made this team not work?
8) Key question: Teamwork/leadership

a. One of the elements mentioned before was leadership. Think back to somebody you've seen work who you think was a very good leader. What did they do or say that made them such a good leader?
b. Can you think of somebody you've seen who was a bad leader? What did they do or say that made you think that?

~~9) Ending question~~

~~a. I'll give you a minute to think about it and then I'm going to go round the group and ask each one of you to tell me what the most important thing a good leader does. So complete the sentence: "A good leader…"~~ (Removed after first focus group. Awkward, seemed repetitive and not in keeping with focus group ethos)

~~10)Ending question~~

~~a. I'll give you a minute to think about it and then I'm going to go round the group and ask each one of you to tell me what the most important thing a good teamworker does. So complete the sentence: "A good teamworker…"~~ (Removed after first focus group)

11)Final questions (leave at least 10-15 minutes for this question to be discussed)

a. Give brief overview of discussion then ask: "Does that sound right?"
b. Then ask: "Is there anything else you want to talk about with regard to professionalism?"

Additional questions:

What are your thoughts on unprofessional behaviour among medical students?

Do you think there are grades of acceptable behaviour, e.g. lying/cheating is acceptable but murder isn't?

What about professionalism in nurses? Would you feel comfortable challenging a nurse's unprofessional behaviour?

Appendix 3-6: Initial email invitation

Dear [First Name],

My name is Michael Moneypenny. I am a specialist registrar in Anaesthesia who is undertaking a clinical fellowship at the School of Medical Education here in Liverpool. I am writing to you because you have been randomly selected to take part in a **research project** which I am conducting. It will only take a total of about **2 and a half hours** of your time **over the next year** or so. One and a half hours will be spent as part of a small **informal group discussion** of 4th year undergraduates about what professionalism means to you (with lunch provided.) The other hour will be spent at the Cheshire and Merseyside Simulation Centre at Aintree Hospital, where you will be able to **lead a scenario on a high-fidelity mannequin** after which we will chat about your decisions and actions. We will provide you with constructive feedback on the scenario and none of your peers will be present during the scenario or chat, so you need not worry about what they will think.

Although the focus group and simulation will be recorded, I will be the only person who will have access to the tapes and I have no involvement at all in grading medical undergraduates. All data will be anonymised and you will not be identifiable in any reports or publications. Nothing you say or do in the focus groups or simulator will be shared with your supervisors/tutors/assessors and you can withdraw from the study at any time without providing a reason and with no consequences to yourself.

As someone who was a medical student only six years ago, I understand the demands placed on your time and I will make sure that the sessions do not clash with your exams or revision time and, as an added incentive, everybody who takes part (approximately 30 undergraduates) will be placed into a draw to **win a new iPod nano**.

Thank you very much for reading this lengthy email and I hope that it has explained a bit about the research project and has gone some way to allay any fears/suspicions about what you will be involved in.
If you agree to take part please email me back and I will send you a consent form which explains more about the project. Please email or call me with any questions you may have.

I very much look forward to hearing from you,
Yours sincerely,
Michael.

Clinical Research Fellow
Centre for Excellence in Teaching and Learning
School of Medical Education
Tel: 0151 794 8379
Email: m.moneypenny@liverpool.ac.uk

Appendix 3-7: Information sheet

Centre for Excellence in Developing Professionalism

School of Medical Education, Cedar House, Ashton Street, Liverpool L69 3GE

30th September, 2008.

EXAMINING PROFESSIONALISM IN MEDICAL UNDERGRADUATES

Participant Information Sheet [Version 1]

You are being invited to participate in a research study. Before you decide whether to participate, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and feel free to ask us if you would like more information or if there is anything that you do not understand. Please also feel free to discuss this with your friends, relatives and GP if you wish. We would like to stress that you do not have to accept this invitation and should only agree to take part if you want to.

Thank you for reading this.

The project:

We are undertaking a research project designed to examine what 'Professionalism' means to undergraduate medical students at the University of Liverpool. The study will use several methods to explore this with a special emphasis towards leadership and team working in medicine.

You have been selected in a random way to participate in this study from your peer group. You will initially be asked to participate in a focus group that will use facilitated discussions to shed more light on this subject.

Later on in the year, you may be asked to participate in a clinical scenario at the Cheshire & Merseyside Simulation Centre. A realistic clinical area with an advanced robotic manikin will be used to assist in the creation of this scenario, which will be appropriate to your level of experience. Your participation will be anonymous and confidential. The scenario will be video recorded and analysed for content.

After the scenario is complete, we will carry out a free-form interview process that will consist of recording your own reflections about the scenario that you have participated in. This will be analysed later for content too. This process will be confidential and all data will be anonymised.

Thus, there will be no way of identifying an individual from the data records in the future. The data will be stored securely, and will only be used for this project.

We believe that the work will enable us to gain an insight into what the undergraduate of the present is thinking with regards to professionalism, and may help us to amend and improve on this element of the curriculum in the future.

Withdrawal of participation

Your participation is voluntary and you are free to withdraw at anytime without explanation and without incurring a disadvantage.

Risks and arrangements

We do not anticipate any risk to you during your participation, but should you experience any discomfort or disadvantage as part of the research then you should make the researcher(s) aware immediately.

If you are unhappy during your participation, or if there is a problem, please feel free to let us know by contacting Dr. Helen O'Sullivan, Director, CETL [0151 795 4356] and we will try to

help. If you remain unhappy or have a complaint which you feel you cannot come to us with then you should contact the Research Governance Officer on 0151 794 8290 (ethics@liv.ac.uk). When contacting the Research Governance Officer, please provide details of the name or description of the study (so that it can be identified), the researcher(s) involved, and the details of the complaint you wish to make.

You will also be covered by the usual University research insurance scheme.

Dissemination of results of the study

We anticipate that the results of the study will be published. When this happens, it will appear on the CETL website and can be accessed publicly. You will not be identifiable from the results.

You can get more information or seek further clarification about the project by contacting any of the following:

Dr. Michael Moneypenny, Clinical Research Fellow, CETL [0151 795 4356]
Dr. Arpan Guha, Hon. Senior Lecturer, School of Medical Education [0151 795 4356]
Dr. Helen O'Sullivan, Director, CETL [0151 795 4356]

Appendix 3-8: Consent form

UNIVERSITY OF LIVERPOOL

CONSENT FORM
[version1 dt. 30th September, 2008]

Title of Research Project: EXAMINING PROFESSIONALISM IN MEDICAL UNDERGRADUATES

Researcher(s): Dr. Helen O'Sullivan
Dr. Arpan Guha
Dr. Michael Moneypenny

	Please initial box
1. I confirm that I have read and have understood the information sheet dated [30/9/2008] for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my rights being affected.	
3. I understand that, under the Data Protection Act, I can at any time ask for access to the information I provide and I can also request the destruction of that information if I wish.	
4. I agree to take part in the above study.	

Participant Name	Date	Signature
Name of Person taking consent	Date	Signature
Researcher	Date	Signature

The contact details of lead Researcher (Principal Investigator) are:

Dr. Helen O'Sullivan, Director, CETL, School of Medical Education, Cedar House, Ashton Street, Liverpool.

Phone: 0151 795 4356 e-mail: H.M.Osullivan@liverpool.ac.uk

Appendix 3-9: Leadership discussion mapped to traits

Challenging the process	
Search out challenging opportunities to change, grow, innovate, and improve. Experiment, take risks, and learn from the accompanying mistakes	”And experience being in similar situations and and how to get out of situations” M1 “A consultant doesn't get to a consultant post by doin all the textbook stuff” M1
Inspiring a shared vision	
Envision an uplifting and ennobling future Enlist others in a common vision by appealing to their values, interests, hopes, and dreams	“someone to unite” M “to see the whole thing as a whole” M “good motivator” M12 “inspire and motivate” M10 “inspires the rest of the team to to do whatever job they're doing”M8 “inspiration and sort of encouragement.” M10 “have a drive towards the goal” F5 “Understanding of the team goals” F6 “points you at the right direction.” F2
Enabling others to act	
Foster collaboration by promoting cooperative goals and building trust Strengthen people by giving power away, providing choice, developing competence, assigning critical tasks, and offering visible support.	“allow like the kind of team members to each do their individual roles all joined” M “somone who can bring out the best of everybody.” M “I can look to him if I need any help” ”delegate to achieve that

	goal”M9 “allowing people space to practice their position in that team as well not overbearing” M6 “being available” F6 “give out tasks” M2 “they shouldn't think they're better than everyone else that's just that is their job as part of the team” M3
Modeling the way	
Set the example by behaving in ways that are consistent with shared values. Achieve small wins that promote consistent progress and build commitment.	“somebody you can look up to” M “leads by example” M8
Encouraging the heart	
Recognize individual contributions to the success of every project Celebrate team accomplishments regularly	”don't take credit for things you know other people have done” F6 “So if you get praised for something that you know someone else has done you'll probably say "Well that was so and so" F6 “respect the position of the nurse staff” F6 “... And then you'll do your sales but also she'll say "Well done" if you've done them.” F1

Appendix 3-10: Principles of Professionalism (TD2003)

Good clinical care: Doctors must practise good standards of clinical care, practise within the limits of their competence, and make sure that patients are not put at unnecessary risk.

> "When I thought about it more maybe competency comes in as well…" M4

> "…it's ehm having the expertise and ehm and ehm using that expertise to the sort of the best sort of fit of the situation" M6

> "It's about kinda working within within your limit as well…" M10

Maintaining good medical practice: Doctors must keep up to date with developments in their field and maintain their skills.

> "I mean you can think of all the other things as well but at the end of the day it's "I know what I'm doing". I think that's the most important thing of ()" M11

Relationships with patients: Doctors must develop and maintain successful relationships with their patients.

> "…someone who takes into consideration the dignity of the patient but also you know lets themselves be a bit human around the patient but there's a line I think." F1

> "…the way you act around people like in doctors interacting with patients and having a relationship with patients." M5

> "I think it's about also about respecting the doctor pa-patient relationship 'cos it is very I suppose intimate relationship…" F9

Working with colleagues: Doctors must work effectively with colleagues.

> "I'd say it's respect sort of respect of your colleagues…" M7

> "Also being professional amongst colleagues as well would make teamwork

'n the team work better obviously with the whole team with physios, OTs, nurses and doctors and everything." M13

Teaching and training: If doctors have teaching responsibilities, they must develop the skills, attitudes and practices of a competent teacher.

Probity: Doctors must be honest.

"I think when you said you put a you know portray as being confident ehm yeah you can do that to an extent but patients eh from my experience I guess prefer honesty..." M10

Health: Doctors must not allow their own health or condition to put patients and others at risk.

Appendix 4-1: Post-scenario questionnaire

1) Was the introduction to the sim centre and the mannequin adequate? **[Response process]**

    a. Yes/No

    Please elaborate

2) Do you think that the scenario tested your leadership and team working skills? **[Content validity]**

    a. Yes/No

    Please elaborate

3) Do you think that the scenario and assessment was fair and acceptable to you as a medical student? **[Acceptability]**

    a. Yes/No

    Please elaborate

4) How realistic was the whole scenario? Please place a cross on the line.

Absolutely unrealistic As real as real life

]_____[

5) Was the think-aloud technique acceptable in terms of ease of performance?

    a. Yes/No

    Please elaborate

6) Were you able to remember why you did/said things during the think aloud technique or did you feel like you had to make up things?

    a. Able to remember/Had to make up things/Bit of both

    Please elaborate

7) Did you find the debrief where we discussed your personal teamwork and leadership useful?

    a. Yes/No

    Please elaborate

8) Is there anything about the hour that you think we should change? Could we improve the experience in any way?

9) Some demographic questions:

Age:

Gender: M/F

Is English your first language: Y/N

Have you ever had a significant leadership role: Y/N

If YES, what was this?

Have you ever had a significant teamworker role: Y/N

If YES, what was this?

Have you been an ALS or critical-care type course? Y/N

If YES, what was this?

APPENDIX 4-2: TEAMWORK AND LEADERSHIP BEHAVIOURS FROM LITERATURE REVIEW

(*Categories/Elements* in italics)

Reference	Teamwork	Leadership	Rating tools
(Brett-Fleegler et al., 2008)	Assumes adequate responsibility when in non-leader roles (airway, circulation)	*Leadership* Has professional attitude toward patient Has professional attitude towards team members Assumes leadership of code Assigns roles Utilizes personnel effectively Communicates effectively with team Performs tasks in appropriate sequence/prioritizes well Intermittently summarizes/ maintains global view	72 questions Yes/No
(Brinkman et al., 2006)	Being respectful/Treating staff with respect Good team member Communicate effectively with staff Complete tasks reliably	Encouraging questions Sharing decisions Accept suggestions Effectively plan course of care	10 5-point Likert items

Reference	Teamwork	Leadership	Rating tools
(Christenson et al., 1998)	N/A	Assessment of the patient Immediate priorities Continual assessment Leadership	4 5-point Likert items
(Cooper and Wakelam, 1999)	*Team dynamics* Information transfer (communication skills) Adaptability (within the roles of their profession) Co-ordination Co-operation Initiative Work effort Team spirit and morale	*Leadership* The leader let the team know what was expected of them (through direction and command) The leader demonstrated the use of uniform guidelines The leader displayed a positive attitude The leader decided what should be done The leader decided how things should be done The leader assigned group members to particular tasks The leader made sure that his part in the team was understood by the team members The team leader planned the work to be done The team leader maintained definite standards of	Team dynamics: 7 5-point Likert items Leadership: 9 5-point Likert items

Reference	Teamwork	Leadership	Rating tools
		performance	
(Dannefer et al., 2005)	Consistently well prepared for sessions; presents extra material; supports statements with appropriate references Always demonstrates respect, compassion and empathy Shares information or resources; truly helps others learn; contributes to the group process; able to defer to the group's needs Seeks appropriate responsibility. Consistently identifies tasks and completes them efficiently and thoroughly Presents him / herself consistently to superiors and peers; trustworthy Admits and corrects his or her own	Identifies and solves problems using intelligent interpretation of data Able to explain clearly his or her reasoning process with regard to solving a problem, basic mechanisms, concepts, etc. Takes initiative and provides leadership Asks classmates and professors for feedback and then puts suggestions to good use Seeks to understand others' views	15 5-point Likert items (and "unable to assess" point)

Reference	Teamwork	Leadership	Rating tools
	mistakes; truthful Dress and appearance always appropriate for the situation Behaviour is always appropriate Directs own learning agenda; able to think and work independently I would refer my own family or patients to this future physician or ask this person to be my own doctor		
(DiMatteo and DiNicola, 1981)	Intelligence Common sense Articulateness Verbal communication Politeness Dedication, diligence and professionalism Appreciation of limitations Cooperativeness and compliance	Ability to teach Leadership Medical-scientific knowledge Sensitivity and perceptiveness	13 4-point Likert items

Reference	Teamwork	Leadership	Rating tools
	Kindness, humaneness, compassion and empathy		
(Epstein et al., 2004)	Consistently well prepared for sessions; presents extra material; supports statements with appropriate references Always demonstrates respect, compassion and empathy Shares information or resources; truly helps others learn; contributes to the group process; able to defer to the group's needs Seeks appropriate responsibility. Consistently identifies tasks and completes them efficiently and thoroughly Presents him / herself consistently to superiors and peers; trustworthy	Identifies and solves problems using intelligent interpretation of data Able to explain clearly his or her reasoning process with regard to solving a problem, basic mechanisms, concepts, etc. Takes initiative and provides leadership Asks classmates and professors for feedback and then puts suggestions to good use Seeks to understand others' views	15 5-point Likert items (and "unable to assess" point)

Reference	Teamwork	Leadership	Rating tools
	Admits and corrects his or her own mistakes; truthful Dress and appearance always appropriate for the situation Behaviour is always appropriate Directs own learning agenda; able to think and work independently I would refer my own family or patients to this future physician or ask this person to be my own doctor		
(Fletcher et al., 2003b) ANTS	*Team working* Co-ordinating activities with team members Exchanging information Using authority and assertiveness Assessing capabilities Supporting others	*Decision making* Identifying options Balancing risks and selecting options Re-evaluating *Task management* Planning and preparing	15 4-point Likert items (and one "not observed" point)

Reference	Teamwork	Leadership	Rating tools
		Prioritizing Providing and maintaining standards Identifying and utilizing resources *Situation awareness* Gathering information Recognizing and understanding Anticipating	
(Gilfoyle et al., 2007)	N/A	*Assign roles to team members* Declare yourself to be in charge of the group Assign PALS algorithm to patient's current condition based on gathered information so far Divide algorithm into distinct steps/actions Recognize skill set of each team member Match members skill set with tasks that need to be done Announce role of each team member to whole team	Variable numbers of questions (17 – 30) depending on scenario with Yes/No/Borderline options

Reference	Teamwork	Leadership	Rating tools
		Assess limitations of team members Recognize skill level of each team member Anticipate difficulty of specific task Compare skill level with difficulty of task to conclude if they are equal Formulate a plan to add skill to team if required *Continuously reassess and re-evaluate progress of resuscitation using all available information* Acknowledge response or lack of desired response to intervention Avoid fixation errors Generate list of reasons why desired result isn't seen Examine patient to choose likely reason from list, or delegate team member to examine and report findings back to you	

Reference	Teamwork	Leadership	Rating tools
		Create solution(s) to problem(s) identified Demonstrate use of another algorithm or approach when expected result to an intervention is not happening *Critically evaluate each team member's performance and redirect him or her as needed:* Observe team member performing assigned task Assess effects of actions of team member If performance is inadequate, causing lack of desired response, then redirect team member to improve skill *Display effective communication during performance of resuscitation:* Use calm, clear voice when talking and giving orders State commands clearly and precisely	

Reference	Teamwork	Leadership	Rating tools
		Avoid making statements into ‘‘thin air’’. Direct your orders to a team member by name. Use closed communication loop: Repeat what has just been said to you and verify meaning of ambiguous messages Encourage open exchange of ideas among team members by listening to all ideas and determining what is important to know or act upon Defer dealing with interpersonal conflicts until after the resuscitation is finished, unless it’s interfering with the performance of the team Quickly manages disruptive behaviour if it is affecting overall team performance	
(Kaye and Mancini, 1986)	N/A	Assessment of both patient status and team performance Dysrhythmia recognition Defibrillation Drug therapy	24 questions Yes/No

Reference	Teamwork	Leadership	Rating tools
		Trouble-shooting	
(Kim et al., 2006)	Communication	Leadership Problem solving Situational awareness Resource utilization	5 7-point Likert items
(Mishra et al., 2008) NOTECHS	*Teamwork & cooperation* Team building/maintaining: relaxed / supportive / open / inclusive / polite / friendly / use of humour / does not compete Support of others: helps others / offers assistance / gives feedback Understanding team needs: listens to others / recognises ability of team / condition of others considered / gives personal feedback Conflict solving: keeps calm in conflicts / suggests conflict solutions / concentrates	*Leadership & Management* Leadership: Involves / reflects on suggestions / visible / accessible / inspires / motivates / coaches Maintenance of standards: subscribes to standards / monitors compliance to standards / intervenes if deviation / deviates with team approval / demonstrates desire to achieve high standards Planning and preparation: team participation in planning / plan is shared / understanding confirmed / projects / changes in consultation Workload management: distributes tasks /	16 4-point Likert items

Reference	Teamwork	Leadership	Rating tools
	on what is right	monitors / reviews / tasks are prioritised / allots adequate time / responds to stress Authority & assertiveness: advocates position / values team input / takes control / persistent / appropriate assertiveness *Problem-solving and decision-making:* Definition & diagnosis: Uses all resources / analytical decision making / reviews factors with team Option generation: suggests alternative options / asks for options / reviews outcomes / confirms options Risk assessment: estimates risks / considers risk in terms of team capabilities / estimates patient outcome Outcome review: reviews outcomes / reviews new options / objective, constructive and timely	

Reference	Teamwork	Leadership	Rating tools
		reviews / makes time for review / seeks feedback from others / conducts post treatment review *Situation awareness:* Notice: considers all team elements / asks for or shares information / aware of available of resources / encourages vigilance / checks and reports changes in team / requests reports / updates Understand: knows capabilities / cross-checks above / shares mental models / speaks up when unsure / updates other team members / discusses team constraints Think ahead: identifies future problems / discusses contingencies / anticipates requirements	
(Moorthy et al., 2005)	*Preoperative preparation* Introduction to team members	*Leadership* Adherence to best practice during the procedure	13 5-point Likert items

Reference	Teamwork	Leadership	Rating tools
	Preoperative instrument and equipment check Briefing *Communication and interaction* Instructions to assistant/scrub nurse: clear and polite Awaits acknowledgment from the assistant/scrub nurse Assistance sought from team members Acknowledges help/advice from team members *Vigilance/situation awareness* Monitored patient's parameters throughout the procedure Awareness of anesthetist Actively initiates communication with	Resource utilization, i.e., appropriate task- load distribution and delegation of responsibilities Authority/assertiveness	

Reference	Teamwork	Leadership	Rating tools
	anesthetist		
(Moorthy et al., 2006)	*Communication and interaction* Instructions to assistant/scrub nurse; clear and polite Awaits acknowledgment from the assistant/scrub nurse Assistance sought from team members *Vigilance/situation awareness* Monitored patient's parameters throughout the procedure Awareness of anesthetist Actively initiates communication with anesthetist during crisis periods *Team skills* Maintains a positive rapport with the whole team	*Leadership and management skills* Adherence to best practice during the procedure, e.g. does not permit corner cutting by self or team Time management e.g. appropriate time allocation without being too slow or rushing team members Resource utilization, i.e., appropriate task-load distribution and delegation of responsibilities Authority/assertiveness *Decision-making crisis* Prompt identification of the problem Informed team members; promptly, clearly, and to all team members Outlines strategy/institutes a plan, i.e., asks scrub nurse for suction, instruments, suture material Anticipates potential problems and prepares a contingency plan, e.g., asks anesthetist to order	19 6-point Likert items

Reference	Teamwork	Leadership	Rating tools
	Open to opinions from other team members Acknowledges the contribution made by other team members Supportive of other team members	blood, calls for help Option generation; takes the help of the team (seeks team opinion)	
(Morison and Stewart, 2005)	*Professional roles and teamworking* Demonstrates knowledge and understanding of, and respect for, the roles of different members of the multidisciplinary team Demonstrates ability to work well with different team members Has ensured that all significant aspects of management of the chronic condition have been addressed by a member of the team Does *not* duplicate information provided by a colleague	N/A	13 4-point Likert items

Reference	Teamwork	Leadership	Rating tools
(Orlander et al., 2006)	N/A	Effectively ran works rounds Created a good sense of open communication on our team Directed the attending physician regarding which patients to discuss and visit as a team Focused the attending on relevant issues Advocated for the team effectively with consultants, nurses, and others Overall leadership effectiveness	7 6-point Likert items
(Pawlina et al., 2006)	N/A	Respect Integrity Responsibility Compassion Problem-solving Commitment to excellence Overall professionalism	7 9-point Likert items (and one "unable to assess" point)
(Robins et al., 2008)	Ability to manage conflict Speak up against a power gradient	Demonstrating leadership	Unable to determine from reference

Reference	Teamwork	Leadership	Rating tools
	Advocate for the patient Demonstrating team orientation		
(Sevdalis et al., 2008) Revised NOTECHS	*Cooperation and Team Skills* Maintains positive rapport with whole team Open to opinions from other team members Acknowledges contribution from other team members Supportive of other team members Conflict handling (concentrating on what is right rather than who is right) *Situation awareness and vigilance* Monitored patient parameters throughout procedure Awareness of anesthetist Actively initiates communication with	*Leadership and Managerial Skills* Adherence to best-practice during procedure (e.g., does not permit corner cutting) Time management (e.g., not being too slow or rushing other team members) Resource utilization (e.g., appropriate task load distribution and delegation of responsibilities) Debriefing the team (e.g., provides details and feedback to the team about procedure) Authority and assertiveness *Decision making* Prompt identification of the problem Informed team members promptly and clearly Outlines strategy and institutes a plan (e.g., asks scrub nurse for suction, instruments, suture	22 6-point Likert items (and one "not applicable" point)

Reference	Teamwork	Leadership	Rating tools
	anesthetist during crisis *Communication and interaction* Instructions to assistant clear and polite Waited for acknowledgement from assistant Instructions to scrub nurse clear and polite Waited for acknowledgement from scrub nurse	material) Anticipates potential problems and prepares contingency plan (e.g., ask anesthetist to order blood, call for help) Option generation (e.g., takes help from others, seeks team's opinion)	
(Wilkinson and Frampton, 2003)	*Interpersonal/communication skills* Ability to relate to patients and colleagues Ability to communicate with patients, their families and other professionals *Management of psychosocial aspects of disease*	*Diagnostic skills* Critically assesses information Identifies major issues and makes timely decisions *Patient management skills* Shows wisdom in selecting treatment Adapts management to different circumstances	12 7-point Likert items

Reference	Teamwork	Leadership	Rating tools
	Ability to recognise and/or respond to psychosocial aspects of illness *Respect* Shows personal commitment to honouring the choices and rights of other persons *Responsibility* Accepts responsibility for own actions and decisions	*Management of multiple complex problems* Ability to manage patients with multiple complex problems *Care skills* Ability to treat patients and coordinate care	
(Wright et al., 2009)	*Assertiveness* Confronting ambiguities and conflicts Asking questions when uncertain Maintaining a position when challenged (and appropriate) Making suggestions Stating an opinion on decisions,	Wright et al. classify leadership as a "teamwork" element	

Reference	Teamwork	Leadership	Rating tools
	procedures, or strategies Adaptable when one's own position is proved to be weak *Decision-making* Communicates possible solutions Gathers information to evaluate solutions Communicates consequences of alternatives Cross-checks information sources Selects the best alternative Development of plans Implements the decisions that were made *Leadership* Explains to other team members exactly what is needed from them during the task Listens to the concerns of other team		

Reference	Teamwork	Leadership	Rating tools
	members Provides statements of team direction, strategy, or priorities for the task Sets goals for the team and orients the team toward those goals Provides feedback to other team members regarding his/her performance *Communication* Verifies information prior to taking an action Acknowledges and repeats messages to ensure understanding Uses accurate terminology Makes concise statements with little extraneous information Establishes and uses conventional or standard speech (e.g.,		

Reference	Teamwork	Leadership	Rating tools
	acronyms/shortcuts) Provides unsolicited responses (gives more detail than was asked, when appropriate) *Situation assessment* Situation assessment updates in which team members communicate the current state of the system Identification of problem situations and recognizing the need for action Exchange of information for the prevention of errors Noting deviations in SA between team members Demonstrated awareness (e.g., via verbal communication) of the on-going mission status and the overall goal		

Reference	Teamwork	Leadership	Rating tools
	Integration of information from multiple sources Accurately prioritizing information and actions		
(Youngblood et al., 2008)	N/A	Knowledge of the Environment Anticipation of & Planning for Potential Problems Assumption of Leadership Role Communication with Other Team Members Distribution of Workload/Delegation of Responsibility Attention Allocation Utilization of Information Utilization of Resources Recognition of Limitations/Call for Help Early Enough Professional Behavior/Inter-personal Skills Overall Behavioral Crisis Management Skills	11 5-point Likert items
(Yule et al., 2008)	*Communication and Teamwork* Exchanging information	*Leadership* Setting and maintaining standards	12 4-point Likert items (and one "not

Reference	**Teamwork**	**Leadership**	**Rating tools**
NOTSS	Establishing a shared understanding Co-ordinating team *Situation Awareness* Gathering information Understanding information Projecting and anticipating future state	Supporting others Coping with pressure *Decision making* Considering options Selecting and communicating option Implementing and reviewing decisions	applicable" point)

APPENDIX 4-3: TEAMWORK AND LEADERSHIP BEHAVIOURS FROM OTHER REFERENCES (*Elements* in italics)

Reference	Teamwork	Leadership	Rating tools
(Baker et al., 2005b)	*Team leadership* Facilitate team problem solving Provide performance expectations and acceptable interaction patterns Synchronize and combine individual team member contributions Seek and evaluate information that impacts team functioning Clarify team member roles Engage in preparatory meetings and feed-back sessions with the team *Mutual performance monitoring*	Baker et al. classify team leadership as a "teamwork" skill.	Not provided

Reference	Teamwork	Leadership	Rating tools
	Identifying mistakes and lapses in other team members actions Providing feedback regarding team member actions in order to facilitate self-correction *Backup behaviour* Recognition by potential back-up providers that there is a workload distribution problem in their team Shifting of work responsibilities to under-utilized team members Completion of the whole task or parts of tasks by other team members *Adaptability* Identify cues that a change has occurred, assign meaning to that change, and develop a new plan to deal with the changes		

Reference	Teamwork	Leadership	Rating tools
	Identify opportunities for improvement and innovation for habitual or routine practices Remain vigilant to changes in the internal and external environment of the team *Team/collective orientation* Taking into account alternative solutions provided by teammates and appraising that input to determine what is most correct Increased task involvement, information sharing, strategizing, and participatory goal setting *Shared mental models* Anticipating and predicting each other's needs Identify changes in the team, task, or team-mates and implicitly adjusting strategies as		

Reference	Teamwork	Leadership	Rating tools
	needed *Mutual trust* Information sharing Willingness to admit mistakes and accept feedback *Closed-loop communication* Following up with team members to ensure message was received Acknowledging that a message was received Clarifying with the sender of the message that the message received is the same as the intended message sent.		
(Carlson et al., 2009)	*Workload management* Roles are clearly and effectively delegated across the group	3 leadership "styles": transactional, flexible/dynamic or neither	4 5-point Likert items

Reference	Teamwork	Leadership	Rating tools
	Communication Key actions/findings are verbalized and clear to group members *Prioritizing and reassessing priorities* Identifies and focuses on key goals initially and reassesses as situation evolves *Vigilance* Group keeps eye on the big picture Avoids fixation, manages distractions, and keeps ahead of the situation by anticipating potential problems		
(Cowan and Cloutier, 1988)	Initial assessment Initiation of life-saving procedures at site of injury	N/A	6-point Likert items

Reference	Teamwork	Leadership	Rating tools
(Cruess et al., 2006)	*Reflective skills* Demonstrated awareness of limitations Admitted errors/omissions Solicited feedback Accepted feedback Maintained composure in a difficult situation *Time management* Completed tasks in a reliable fashion *Interprofessional relationship skills* Demonstrated respect for colleagues Assisted a colleague as needed Respected rules and procedures of the system		24 4-point Likert items (and one "not observed/not applicable" point)
(Gaba et al., 1998)	Inquiry/assertion Communication	Orientation Leadership	11 5-point Likert items and 2 5-point

Reference	Teamwork	Leadership	Rating tools
	Feedback Group climate Workload distribution Vigilance	Anticipation/planning Vigilance Re-evaluation	global ratings (for crew and for leader)
(General Medical Council and Medical Schools Council, 2009)	Recognise and work within the limits of their competence and ask for help when necessary Be able to work effectively in a team and to take on different roles as appropriate, including taking responsibility for tasks Develop and demonstrate teamwork and leadership skills Be aware of the roles and responsibilities of other people involved in delivering healthcare Raise concerns about overall practice in a healthcare setting or about colleagues, including other students, medical	Develop and demonstrate leadership skills	N/A

Reference	Teamwork	Leadership	Rating tools
	practitioners and other healthcare workers, with the appropriate person if patients are at risk of harm.		
(General Medical Council, 2009)	*Learn and work effectively within a multi-professional team* Understand and respect the roles and expertise of health and social care professionals in the context of working and learning as a multi-professional team Understand the contribution that effective interdisciplinary teamworking makes to the delivery of safe and high-quality care Work with colleagues in ways that best serve the interests of patients, passing on information and handing over care, demonstrating flexibility, adaptability and a problem-solving approach Demonstrate ability to build team capacity	Use their ability to provide leadership Expected to offer leadership, and to work with others to change systems when it is necessary for the benefit of patients	N/A

Reference	Teamwork	Leadership	Rating tools
	and positive working relationships and undertake various team roles including leadership and the ability to accept leadership by others		
(Holcomb et al., 2002)	Other members assume functional roles Verbal communication within team Systematic and orderly assessment Ability to handle distractions	Clearly defined team leader emerges	46 3-point Likert items
(Howard et al., 1999)	N/A	Competence Vision Team leadership Planning skills Persistence Implementation skills	N/A
(Hughes et al., 2008)	N/A	*Regular attendance at group meetings* Attended all or almost all meetings, stayed to agreed end, worked within timescale, active and attentive, prepared to be flexible about meeting	Free text

Reference	Teamwork	Leadership	Rating tools
		times. *Contribution of ideas to the task* Usually thought about the topic in advance of the meeting, provided workable ideas which were taken up by the group, built on others' suggestions, and was prepared to test out ideas on the group rather than keep quiet. *Researching, analysing and preparing material for the task* Did what they agreed to do, brought materials, did an adequate share of the research and helped to analyse and evaluate the material. *Contribution to cooperative group process* Left personal differences outside the group, willing to review group; progress and tackle	

Reference	Teamwork	Leadership	Rating tools
		conflict in the group, took on different roles as needed, kept group on track, willing and flexible but focused on the task. *Supporting and encouraging group process* Listened to others, encouraged participation, enabled a collaborative learning environment, sensitive to issues affecting group members, supported group members with special needs. *Practical contribution to end product* Willing to try new things. Did not hog the tasks, made a high level of contribution, took own initiative, was reliable and produced good standard work/presentation.	
(Ker et al., 2003)	Collaborative teamworking Ability of the team to prioritise the workload	Effective leadership	Free text

Reference	Teamwork	Leadership	Rating tools
	Competence in clinical performance		
(Malec et al., 2007)	Each team member demonstrates a clear understanding of his or her role The team prompts each other to attend to all significant clinical indicators throughout the procedure/intervention When team members are actively involved with the patient, they verbalize their activities aloud Team members repeat back or paraphrase instructions and clarifications to indicate that they heard them correctly Team members refer to established protocols and checklists for the procedure/intervention All members of the team are appropriately involved and participate in the activity Disagreements or conflicts among team	A leader is clearly recognized by all team members The team leader assures maintenance of an appropriate balance between command authority and team member participation	16 3-point Likert items (and one "not observed point")

Reference	Teamwork	Leadership	Rating tools
	members are addressed without a loss of situation awareness When appropriate, roles are shifted to address urgent or emergent events When directions are unclear, team members acknowledge their lack of understanding and ask for repetition and clarification Team members acknowledge—in a positive manner—statements directed at avoiding or containing errors or seeking clarification Team members call attention to actions that they feel could cause errors or complications Team members respond to potential errors or complications with procedures that avoid the error or complication When statements directed at avoiding or containing errors or complications do not		

Reference	Teamwork	Leadership	Rating tools
	elicit a response to avoid or contain the error, team members persist in seeking a response Team members ask each other for assistance prior to or during periods of task overload		
(Mickan and Rodger, 2005)	Communication Cohesion Mutual respect	Purpose Leadership Goals	
(Mitchell and Flin, 2008)	Communication Teamwork	Situation awareness	N/A
(Morgan et al., 2007)	Obstetricians gave feedback to the anesthesiologist Anesthesiologists gave feedback to the obstetricians Physicians gave feedback to the nurses	The obstetrician/anesthesiologist encouraged questions from the obstetric resident The successful management of the scenario was mainly a function of the obstetrician's/anesthesiologist's expertise	45 5-point Likert items One global rating 5-point Likert of team performance

Reference	Teamwork	Leadership	Rating tools
	Nurses gave feedback to the physicians. The anesthesiologist/obstetrician/nurses took charge of coordinating the team effort The obstetrician took charge of coordinating the team effort The team effectively prioritized activities Conflicts were openly resolved The team worked well together	The obstetric resident should have been more involved in the patient's care The anesthesia resident should have been more involved in the patient's care The successful management of the case was mainly due to the technical proficiency of the physicians During the critical event management, the nurses were appropriately consulted by the physicians The nurses assumed a leadership role during the scenario.	
(Ottestad et al., 2007)	Communication (words leading to action) Information transfer Communication content Information use	Anticipation and planning Leadership Task distribution	7 5-point Likert items
(Rodgers et al., 2009)	N/A	The team leader assured that high-quality CPR was in progress The team leader assigned team member roles	14 7-point Likert items

Reference	Teamwork	Leadership	Rating tools
		The team leader assured that monitor leads were applied appropriately The team leader assured the airway was being managed appropriately The team leader recognized the initial ECG rhythm The team leader properly utilized defibrillation The team leader ordered the correct medication treatment for the initial rhythm The team leader followed the appropriate ACLS algorithm The team leader recognized the ECG rhythm changes The team leader provided appropriate post arrest care The team leader demonstrated confidence The team leader appeared knowledgeable	
(ten Cate and de	Courteousness and respect Adequate information giving	Adequate information gathering	N/A

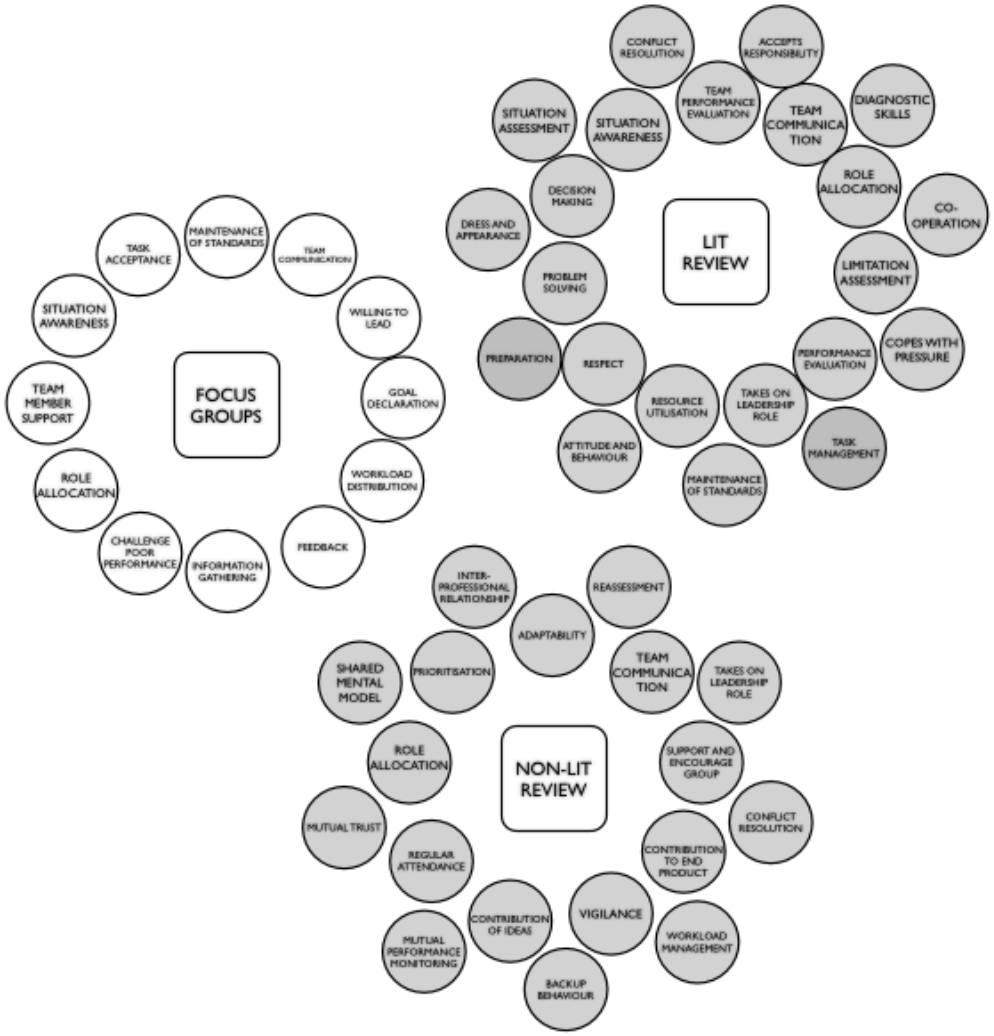
Reference	Teamwork	Leadership	Rating tools
Haes, 2000)	Handling emotions; empathy Structuring communication Insight into one's own emotions, norms, values and prejudices Adequate cooperation with nurses and colleagues Knowing one's own limits, willingness to critically assess one's own behavior, adequate handling of feedback Display of dedication, sense of responsibility and engagement		
(Undre et al., 2007a)	Maintains a positive rapport with the whole team Open to opinions from other team members Acknowledges the contribution made by other team members Supportive of other team members	Adherence to best practise during the procedure; e.g., does not permit corner cutting by self or team Time management; e.g., appropriate time allocation without being too slow or rushing team members Resource utilization; i.e., appropriate task-load	23 6-point Likert items (and one "not applicable" point)

Reference	Teamwork	Leadership	Rating tools
	Conflict handling; e.g., concentrates on what is right rather than who is right	distribution and delegation of responsibilities Debriefing the team; i.e., provides details and feedback to the entire team about the procedure Authority/assertiveness	
(Undre et al., 2007b) OTAS	Communication Coordination Cooperation/backup behavior Monitoring/awareness	Leadership	5 7-point Likert items
(Varkey et al., 2009)	Teamwork skills	*Emotional intelligence* Self-awareness Empathy Cultural sensitivity Professionalism Drive Inspiration Commitment Appropriate balance of confidence and humility	12 4-point Likert items

Reference	Teamwork	Leadership	Rating tools
		Communication skills Listening and incorporating others' views Articulating a vision *Management skills* Conflict resolution Delegating Organization Time management Decision-making Negotiation Quality improvement skills	
(Wagner et al., 2009)	Interpersonal communication	N/A	Unable to determine from reference
(Wallin et	*Teamwork competencies*	N/A	11 5-point Likert

Reference	Teamwork	Leadership	Rating tools
al., 2007)	Knowledge of the environment Anticipation of and planning for potential problems Assumption of leadership role Communication with other team members Distribution of workload / delegation of responsibility Attention allocation Utilisation of information Utilisation of resources Recognition of limitations / call for help early enough Professional behaviour / Interpersonal skills Overall team leadership skills		items

Appendix 4-4: Loci



Appendix 4-5: Triangulation of 3 sources into final assessment tool

FOCUS GROUPS	LIT REVIEW	NON LIT REVIEW	FINAL ASSESSMENT TOOL
Challenge poor performance	Performance evaluation	Mutual performance monitoring	Challenges leader if appropriate
	Team performance evaluation		
	Limitation assessment		
Feedback		Contribution of ideas	Provides appropriate feedback to team leader and team-members
Goal declaration	Decision making	Prioritisation	Declares goal and how to achieve it, changing this if necessary as new information is collected
		Shared mental model	
Information gathering	Situation assessment	Reassessment	Solicits opinions from team-members
	Diagnostic skills		
Maintenance of standards	Maintenance of standards		Not included
Role allocation	Role allocation	Role allocation	Allocates roles/tasks to appropriate team members and ensures workload is shared
Situation awareness	Situation awareness	Vigilance	Maintains situational awareness or ensures SA maintained by another if leader distracted
Task acceptance	Accepts responsibility	Regular attendance	Accepts and completes task
Team communication	Team communication	Team communication	Listens to team-members and responds appropriately
Team member support	Co-operation	Support and encourage group	Supports other team members
		Inter-professional relationship	
	Respect	Backup behaviour	

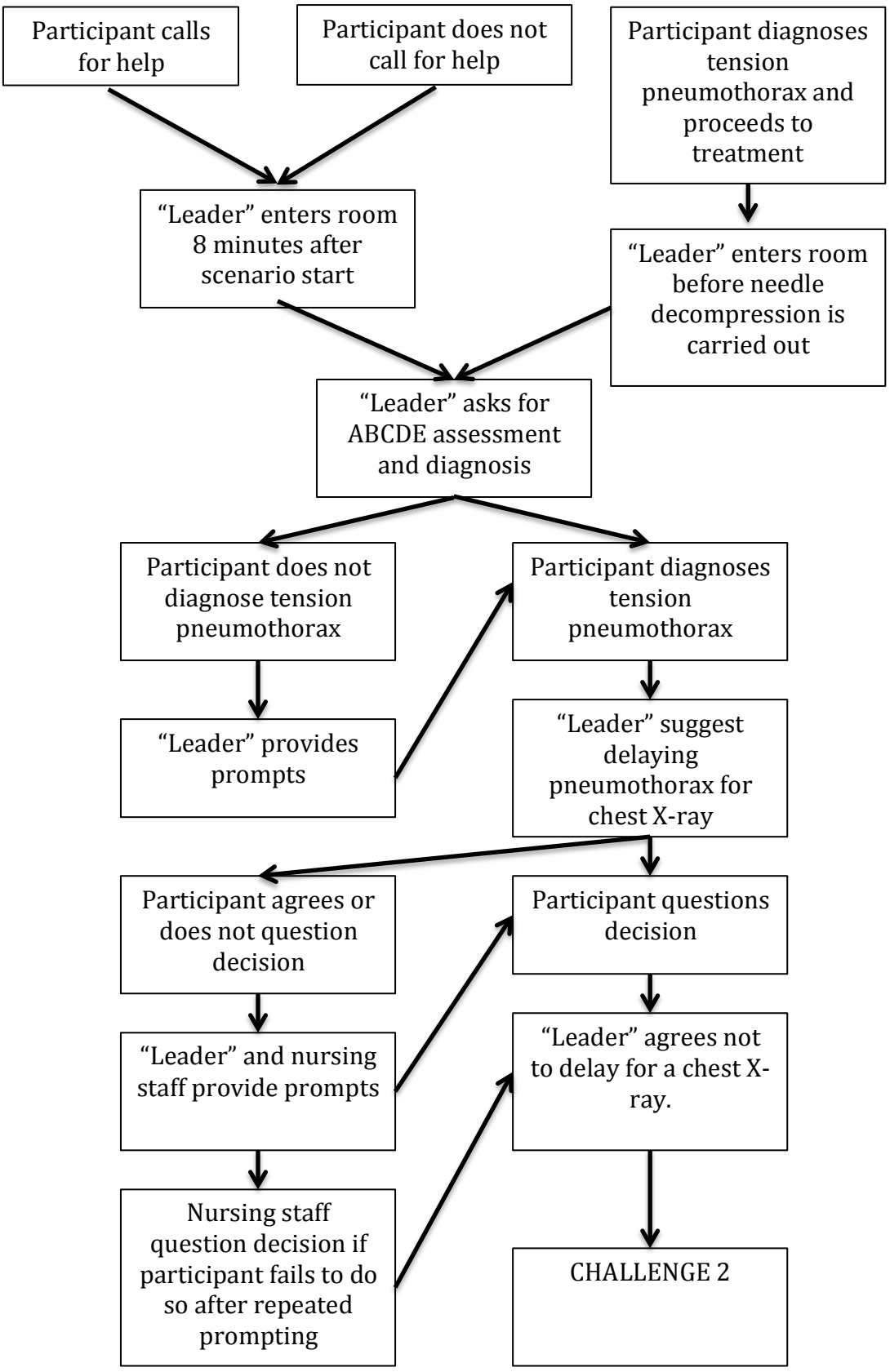
FOCUS GROUPS	LIT REVIEW	NON LIT REVIEW	FINAL ASSESSMENT TOOL
		Mutual trust	
Willing to lead	Takes on leadership role	Takes on leadership role	Adopts leadership role if necessary
Workload distribution	Resource utilisation	Workload management	Allocates roles/tasks to appropriate team members and ensures workload is shared
	Attitude and behaviour		Supports other team members
	Conflict resolution	Conflict resolution	Challenges leader if appropriate
	Copes with pressure		Not included
	Dress and appearance		Not included
	Preparation		Not included
	Problem solving	Adaptability	Declares goal and how to achieve it, changing this if necessary as new information is collected
	Task management	Contribution to end product	Accepts and completes task

Appendix 4-6: Assessment tool

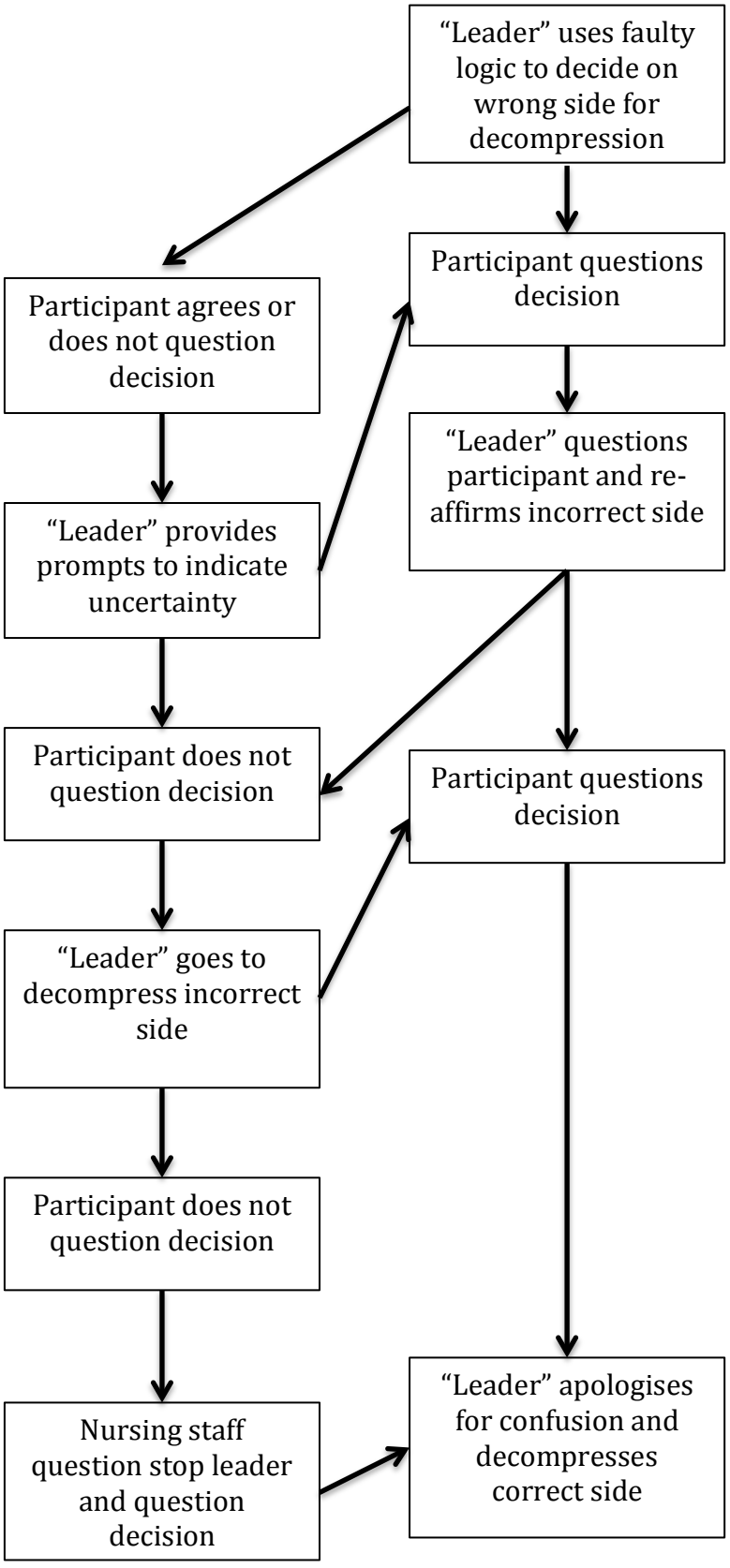
TEAM WORKING	VERY POOR	POOR	ACCEPTABLE	GOOD	VERY GOOD	UNABLE TO ASSESS
1) Accepts and completes tasks						
2) Provides appropriate feedback to team leader and team-members						
3) Adopts leadership role if necessary						
4) Supports other team-members						
5) Challenges leader if appropriate						
GLOBAL SCORE 1						
GLOBAL SCORE 2	UNACCEPTABLE		ACCEPTABLE			
LEADERSHIP	VERY POOR	POOR	ACCEPTABLE	GOOD	VERY GOOD	
1) Listens to team-members and responds appropriately						
2) Allocates roles/tasks to appropriate team-members and ensures workload is shared						
3) Declares goal and how to achieve it,						

changing this if necessary as new information is collected.						
4) Maintains situational awareness or ensures SA maintained by another if leader distracted						
5) Solicits opinions from team-members						
GLOBAL SCORE 1						
GLOBAL SCORE 2	UNACCEPTABLE		ACCEPTABLE			

Appendix 5-1A: Challenge 1



Appendix 5-1B: Challenge 2



Appendix 5-2: Did participants know that the leader's decision was wrong?

Potential challenge	Candidate aware?	Think-aloud transcript
1	Y	And then I think this guy just wants to chill and sit here and I just that's not a good idea, we need to do something now
2	Y	I'm thinking "Hang on, this isn't right." Wait no, don't do it. I'm thinking hang on a minute
3	Y	Obviously an x-ray's important but it's an emergency situation so as soon as he said x-ray I was thinking: "What what what're we doin here? This isn't right."
4	Y	I wasn't too happy about when he said that because I was like well if he's got no poor air entry the little I do remember from Medicine last year (laughs) actually I kind of remember that you know could be a pneumothorax and obviously when he mentioned that the trachea was deviated as well that's it just didn't seem like the right call at the time. So I thought maybe not not try 'n attack him cos obviously he's my senior
5	Y	And then I felt 20 minutes was too long to wait for a chest x-ray in this situation. So then I kinda thought he was wrong (laughs)
6	Y	…because then we talked about the chest x-ray which was something that I knew we needed to do but then it dawned on me that you'd never wait for a chest x-ray when you suspect a pneumothorax
7	Y	I learnt this like this for my exams and I was quite happy it was on the other side

8	Y	Even though I've been told that () in my in my own knowledge is that we shouldn't wait for a chest x-ray so cos he's more senior than me I shouldn't argue with him. So I I knew eh from the clinical signs it indicated the diagnosis but because he's more… he had greater knowledge than me I was prepared to listen to him.
9	Y	I knew that wasn't right but I wasn't… (laughs) wasn't sure sure whether I should intervene or not
10	Y	I was thinkin what the hell is he on about… Yeah 20 minutes I was that's (laughs) definitely not right but then again I'm still thinking he's he's obviously senior isn't he so ehm he obviously knows…
11	Y	At that point I was like remembering that we should have stick a needle in before ordering chest x-ray
12	Y	Ehm I stepped back then I was quite happy to have a more senior doctor helpin me out at this stage
13	Y	So I'm think.. I just went along with her there even though I knew she was wrong.
14	Y	I was like that doesn't make any sense there's no hole on that side (laughs) I was thinking there is only a hole on the left side surely it can't be the right side… I shifted my ground (laughs) I knew it but I shifted my ground really
15	Y	And being an F1 I just wasn't sure whether I should address my concern properly or not
16	?	But then he kept talking and I was like oh oh ok and then he started talking about the fluid and I was like oh ok, you're in charge but I did want to go and get the thing to start with..